Master-thesis

Performance evaluation of quality metrics for video quality assessment

Fakultät IV - Elektrotechnik und Informatik
Assessment of IP-based Applications (AIPA)

Samuel Abeijón Monjas
September 15, 2010

Examinator: Alexander Raake, Ph. D.

Tutor: Alexander Raake, Ph. D.

Acknowledgements: I thank Sarah Correa for all these years of affection and support. Ernest Granados, Alex González and friends for the great moments in the university. The Erasmus and not Erasmus students for the great time in Berlin, especially: Loreto, Javi, Conchi and Jorge. Lorena Hernández, "un amigo", for the last month in Berlin and for her wholehearted support. This theis is fundamentally dedicated to my father, mother, grandmother and brother for being my family. Specially for my grandfather, who couldn't see the end.

Berlin, September 15, 2010 Samuel Abeijón Monjas

List of Figures

List of Tables

Contents

1 Introduction

1.1 Video quality assessment

The proliferation of live-video services like real-time streaming in the Internet or the so called triple play services which provide IPTV possibilities, has made the quality limitations more and more relevant for users and providers. Research on objective quality assessment of video streams is still in a early phase and a lot of work has to be done.

Since most of internet bandwidth is consumed by video [1] some algorithms which model the subjective quality perceived by the user need to be developed. Actually the research for voice streaming is quite mature (ITU E-Model [2]) but the question of how to model the user's perception of video is still open. Despite the fact that many algorithms have been proposed lately for objective video quality assessment, their performance is still not satisfactory.

Quality of Experience (QoE) has become a term commonly used to describe the application- and user-oriented quality of video and multimedia services. Defining quality of experience is a hard task since it depends on the opinion of each single person, the human perception. In [3] some of the numerous factors contributing to QoE are listed:

- **Factor of attention** depends on the interest of the viewer. Viewers do not have the same level and focus of attention in what they consider an uninteresting program as in one of their favorites programs.
- **Quality expectations** of the viewer are not always the same. They have different expectations if watching a film in a cinema or a short clip on a computer.
- **Video experience** of the viewer. If the viewer knows about a better technology he will be able to expect a better quality. This determines quality expectations.
- **Display type and properties** where the sequence is screened. The quality experience is different if the user watches the video sequence in a LED screen than in a catodic tube monitor and also the size, resolution, color, etc. determines the quality. The viewer will expect better quality in better display devices.
- **Viewing setup and conditions** where the film is being watched: distance to the screen, light, etc.
- **Quality and synchronization of the audio** will affect the perceived quality. If the image is perfect but the audio is not well synchronized the experience of the viewer will be very poor.
- **Interaction whit the service or display** like the remote control, program guide, etc.

As all these factors mentioned above can not be taken into account together, because of the wide variety and subjectivity, most of quality metrics only account for some of them and focus on measuring fidelity between original and distorted videos. However, two challenging issues remain:

- Video systems are complex and consist of lots of components which process the video in some way and might affect its quality.
- Visual perception is even more complex. It is needed to understand how people perceive video and its quality, and once again the field of the subjectivity is entered.

Since digital video data, which are stored in video databases and distributed through communcations networks, are subject to various kinds of distortions during acquisition, compression, processing, transmission, decoding and reproduction, it has become an important issue for a video service system to be able to quantify the video quality degradations that occur in the system. The target is to mantain control and possibly enhance the quality of the video data. It is crucial to find an effective image or video quality assessment metric for this purpose.

Three typical situations where quality degradation occurs are presented below:

- Lossy video compression techniques, which are usually used to reduce the required bandwidth for the video data transmission, may degrade the quality during the quatization process.
- Digital video bitstreams transmited over error-prone channes, like wireless channels, may be received imperfectly because of the impairment occurred during the transmission.
- Internet is a package-switched communication network, which can cause loss or severe delay of data packages, depending on the network conditions and the quality of services.

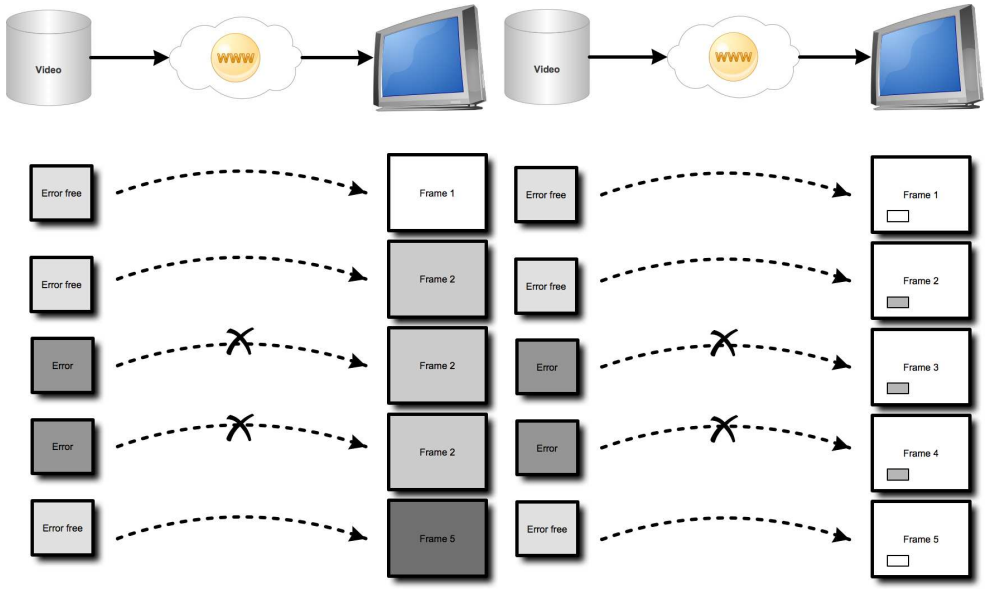
Since people are the ultimate receivers in most of the video applications, the most reliable way of assessing the quality of the videos is to make a subjective evaluation. The mean opinion score (MOS) is a subjective quality measurement of evaluations made by people. These scores are then used to evaluate objective metrics with the objective of knowing which of them correlate better with the subjective results.

The main target applications for an objective image or video quality assessment metric are described in the following three points:

1. Monitoring of image or video quality for quality control systems. For instance, a video acquisition system can check the quality between different providers and record the one with the best quality by automatically monitoring and adjusting itself to obtain the best image or video quality. A company which provides video services can examine the quality of the transmitted video on the network to control the video steaming and offer a quality of service (QoS).
2. If multiple video processing systems are available it can be employed to benchmark image or video processing systems and algorithms. The video quality assessment metric will help in the task of determining which of them provides the best quality of experience.
3. If video transmission or processing systems are variable it can be embedded into them and be used to find the best parameter settings of the algorithm or the best compression rate. For example, in a communication system the image or video quality assessement metric can help in the optimal design of the prefiltering and bit assignment algorithms at the enconder and the optimal reconstruction, error concealment and postfiltering algorithms at the decoder.

In non real-time applications, such as Email, if a transmission error occurs during the transmission then the application would ask for a retransmission till the whole content is free of errors. In IPTV, which is an Internet television service where the contents are delivered using the architecture and networking methods of the Internet Protocol Suite

over a packet-switched network infrastructure, the packets just come one after the other and there is no time for asking for a retransmission. To conceal these type of errors different methods exist that will be taken into account in this thesis: the freezing and the slicing method. Packet loss concealment is a technique to mask the effects of packet loss in communications.



(a) Freezing (b) Slicing

Figure 1: Operation of the freezing and slicing methodologies.

Freezing: as depicted in Figure 1a, the last error-free frame will be repeated till the next complete and error-free frame arrives. The viewer will have the impression that the video was paused for a short time. In the figure each frame is represented with one color. When the first lost packet is detected the color of the image doesn't vary, it stays the same as the previous one. When the following error-free packet arrives, the corresponding image is shown.

Slicing: as depicted in Figure 1b, each frame is divided in small areas called slices. In case one of the packets corresponding to one of the areas is missing it can be possible that the user doesn't notice because of the movement of the whole frame. After the first error-packet is detected the sequence continues showing the next frames with an error area. The image show in this area corresponds to the last error-free packet corresponding to the distorted area. The missing area is concealed by the neighbouring pixels. When the following error-free packet arrives the sequence continues with the original frame. When using slicing methods macroblocks artifacts appear and the question of how this affects to the user becomes even more difficult to answer.

However, other concealment methods may be adopted by the decoder.

1.2 Thesis outline

This thesis will focus on the performance evaluation of existing metrics designed for predicting the image or video quality, respectively. Selected metrics that are not already

implemented will have to be implemented and will be applied to existing databases of video sequences. Their performance will be compared with ratings of subjects that have been obtained in previous empirical tests. Moreover, a novel method for objective video quality assessment will be presented. This method improves the correlation of objective results with the subjective ratings and exhibits robust behaviour across all tested scenarios.

In section 2 of this thesis the most important algorithms for video quality assessment will be presented. Firstly, the simplest and most widely used full-reference algorithms, Peak-Signal-to-Noise-Ratio (PSNR), computed by averaging the square intensity differences of distorted and reference video. It is appealing because it is simple to calculate, has clear physical meanings and is mathematically convenient in the context of optimization, although it is not very well matched to perceived visual quality. Secondly, the structural similarity (SSIM), which is an algorithm based on the human visual system (HVS) and tries to take advantage of it by taking into account the hypothesis that the HVS is highly adapted for extracting structural information. Thirdly the video quality metric (VQM) algorithm, which is a reduced reference metric that combines different parameters using linear models to produce estimates of video quality that closely approximate subjective test results. Moreover, the macroblock concealment efficiency (MCE) algorithm, which is a no reference model that based on the number of macroblocks containing errors, which have not been possible to be concealed. VQuad is another full reference quality metric that uses perceptual degradation measures like blockiness, tilling, blurring, jerkiness and perceptual differences to estimate a MOS prediction. And, finally, SSIM+, which is a full-reference metric based on the original SSIM that improves the correlation with the subjective results by proposing a novel way to extract structural information from the scenes.

In section 3 the results of the tests on two different databases for all the mentioned algorithms are given. The TLabs database, which has two phases: 1+ and 1++, consists of five different videos contents with different characteristics and distortions, and the LIVE database consists of ten different contents with several kinds of distortions as well. The limitations of some all the algorithms which are based on the human visual system are also analitically discussed in section 3.

Section 4 provides a summary of all the conclusions taken from the tests. Some proposals of future work are described in this section, too.

2 Video quality assessment algorithms

2.1 Mean of Opinion Scores measurement methodology

Subjective quality assessment has been an essential research topic for a long time. Objective video quality assessment methods require subjective experiments, where the quality of each tested video is produced by the mean of opinion scores (MOS). These subjective quality assessments are the most accurate and reliable way to evaluate the perceptual video quality and provide the "ground truth" for the evaluation and comparison of both methodologies.

Based on several inputs of Video Quality Experts Group (VQEG) the International Telecommunication Union (ITU) suggested detailed protocols to standarise subjective quality evaluation tests. In 2002, the first recommendation, ITU-R Rec. BT.500, was published for TV and in 2008, ITU-T Rec. P910, for multimedia. Both of them are similar and suggest the most commonly used procedures like the viewing environment, the criteria for the selection of the viewers, tests, videos, data, etc.

The suggested methods are categorized in two groups, double and single stimlus. The following are brief descriptions of the most common test schemes. The interested reader may refer to [4, 5, 6] for more details.

- **Double-stimulus Methods:**

  - Double-Stimulus Continuous Quality Scale (DSCQS)
    In this scheme two videos are presented twice to the viewer, the source unimpaired sequence and the test processed version of the sequence. The order is randomized and at the second presentation of each video the viewer has to make his rating, see Figure 2. For the rating the viewer uses a grading scale like the one shown in Figure 3. VQEG estimates that DSCQS is the most reliable method but its deficiency relies on the redundancy of the ratings limited by the scale of the test sequences.

A Source or Processed 8 s	gray 2 s	B Processed or Source 8 s	gray 2 s	A' Source or Processed 8 s	gray 2 s	B' Processed or Source 8 s	gray 6 s

Figure 2: Presentation structure of test material for DSCQS (from ITU-T Rec. BT.500).

  - Simultaneus Double-Stimulus for Continuous Evaluation (SDSCE)
    Since the duration of previous methods was limited to 10 seconds and therefore not representative of much longer videos happening in real service, a long test evaluation method was needed. Thus, SDSCE tests sequences last, at least, 2 minutes. In this case the viewer is required to watch two video sequences, the source distortion-free version and the test processed version, at the same time, see Figure 4. Then they have to check the differences between the two

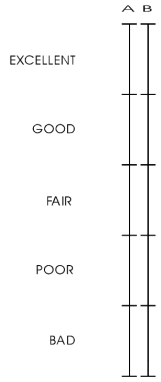


Figure 3: DSCQS grading scale (from ITU-T Rec. BT.500.).

sequences. The problem on that methodology is that the viewers have to shift their attention between two pictures.

Figure 4: Example of display format for SDSCE (from ITU-T Rec. BT.500).

- **Single-stimulus Methods:**

  – Absolute Category Rating (ACR)
  This method tests the subjective quality without any explicit reference. It is very efficient since in a short time a large number of sequences can be tested. The references are assumed to be perfect although in some cases in the capture phase some artifacts are introduced, owing to that fact, ACR-Hidden Reference (ACR-HR) was developed. Here the original version is inserted randomly in the test dataset. Then the differential MOS (DMOS) between the reference and the test sequence may be calculated.

  – Single-Stimulus Continuous Quality Evaluation (SSCQE)
  SSCQE was designed for measuring continuous subjective quality of longer video sequences, in the order of 30 minutes long. The viewer is asked to give quality rating instantaneously with a sliding bar as the video is playing. Since some delays between the viewers can be given the scores need to be calibrated rather than simply averaged over the time.

The main difference between the Double-stimulus and Single-stimulus methods relies on the capability of the Double-stimulus methods to compare the reference and the processed sequences, therefore it is more precise but, also, requires more time. On the other hand Single-stimulus methods permit more votes to be obtained, thus, increase the accuracy of the tests. In the end it has been concluded that both have similar performances.

When designing the tests schemes some precaution have to taken. The tests do not have to be too complex, or the viewer will have difficulties to understand it and therefore not perform a valid test. Too much time consuming will decrease the vote accuracy and since lots of dedicated sources are needed these type of tests are expensive.

2.2 Categorization of video quality assessment methods

For characterizing the quality of video and predict the viewer rating of the transmited data, objective quality metrics were designed. Different types of objective metrics exist [7]. As depicted in Figure 5,the first classification has to be done between the type of data that is received. In the decoded video analysis two types of metrics can be distinguished: data metrics, which measure the fidelity of the signal without considering its content and picture metrics, which treat the video data considerating the contained visual information. In case of compressed video delivery over packet networks (internet) packet- or bitstream-based metrics are used. They look at the packet header information and the enconded bitstream directly without decoding the video. Another type of classification is possible depending on the amount of reference of information they require: full-reference, no-reference and reduced-reference. These classifications are discussed next [3].

1. **Signal pixel-based video quality models**

   (a) **Data metrics** Data metrics have become quite popular thanks to metrics like MSE or PSNR, which will be explaind in detail later. They are based on a byte-by-byte comparison of the data without considering the type of content that they represent. These metrics ignore pixels and their spatial relationship or how a human would interpretate the different images contents.

   Since they were designed to characterize data fidelity but whithout taking into account the content they represent, a number of lost packets will have the same effect on the measure of the perceived quality. The visual importance of the packets and the bits concerned is ignored by these kind of algorithms.

   - Data metrics are distortion-agnostic. For these kind of metrics it doesn't matter where the distortions take place. Noises that might be more sensitive for the human visual system will be rated the same as noises which wouldn't affect a person.
   - Data metrics are content-agnostic. Depending on the part of the image or video where the distortion takes place the viewer perception varies. Therefore a noise in a part of an image with a lot of image activity from the content itself (edges, texture, etc.) would be masked by the image itself. But in a region devoid of content activity the distortion would be rapidly noticed by the viewer. In data metrics both distortions would be rated the same.

   (b) **Picture metrics** Because of the problems pointed aboved, some new lines of research have been opened. New quality assessment metrics that specifically account for the effects of distortions and content on perceived quality are wanted.

Those can be classified in two groups: the vision modeling approach and the engineering approach [8].

i. The vision modeling approach tries to incorporate aspects of the human visual system (HVS) which can help to define a picture and therefore to do a more accurate video quality assessment. These aspects can be color perception, contrast sensitivity and pattern maskin, which are modeled by extracting data from psychophysical experiments. Due to their generality, these metrics can in principle be used for a wide variety of video distortions.

ii. The engineering approach is based primarily on the extraction and analysis of certain features or artifacts in the video. Like structural elements from the scene like contours, edges, etc. or specific distortions introduced by a particular video processing step (e.g. compression, transmission channel), for instance: block artifacts. To make the estimation of the quality the metrics evaluate the strenght of these features. This metrics take also into account some psychophysical effects like in the previous approach, but this metrics are fundamentally based on the image content and distortion analysis.

2. **Bit-stream based video quality models**

   (a) **Packet- and Bitstream-based metrics** Since video services over IP networks, like Internet streaming or IPTV, are increasing, some metrics base their quality measurement on the impact of network losses. As losses directly affect the encoded bitstream, such metrics are often based on parameters that can be extracted from the transport stream and the bitstream whith no or little decoding. These kind of metrics have the advange that much lower data rates and lower bandwidth is required compared to metrics which look at the fully decoded video. Also they can measure the quality of more than one video stream in parallel. On the other hand, these metrics need to understand the specific codecs and network protocols of each transmission, therefore they need to adapt. These so-called "hybrid" metrics use a combination of packet information, bitstream or even decoded video as input. As they use bitstream information, they need of networks free of packet-loss and, in general, cannot be used to evaluate video quality due to packet loss. This models typically use DCT coefficients and quantizer scales [9, 10]. Another approach is to calculate the number of pixels or macroblocks which cannot be decoded correctly [11].

3. **Amount of reference information**

   According to the availability of the original image or video signal, which is considered to be error-free, the quality assessment metrics can be classified. The reference image or video frame will be compared with the distorted one. In case of a full-reference metric a pixel-by-pixel comparison between both images or frames will be done. When using a no-reference metric the received image or video frame under test will be used to calculate the perceived quality. Reduced-reference metrics extract some features from the original and received video and will use them to determine the QoE.

   A detailed explanation is given below:

   (a) **Full-reference:** A frame-by-frame comparison between a reference video and the video under test is performed. As shown in Figure 6a full-reference (FR)

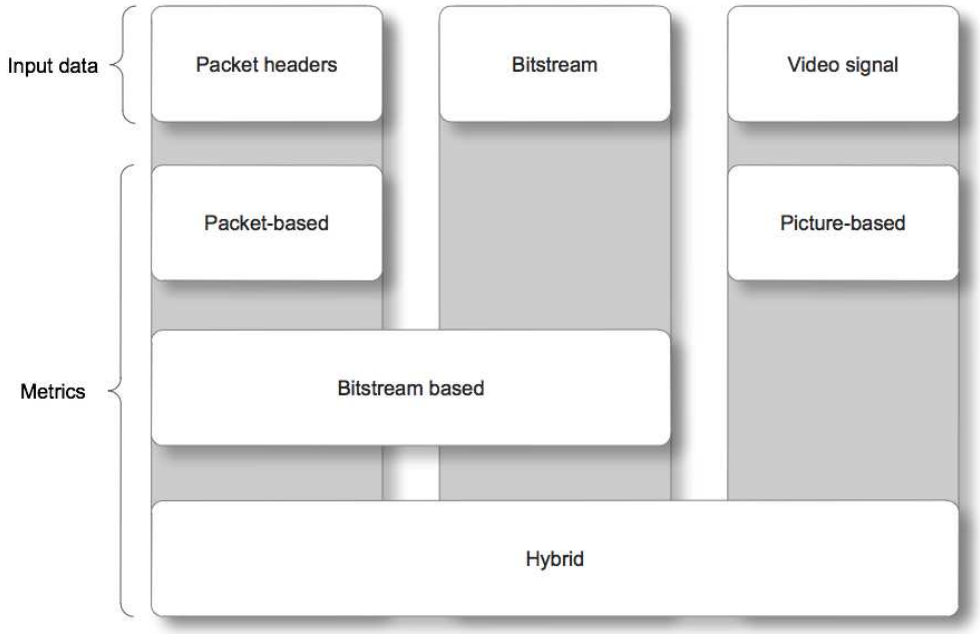


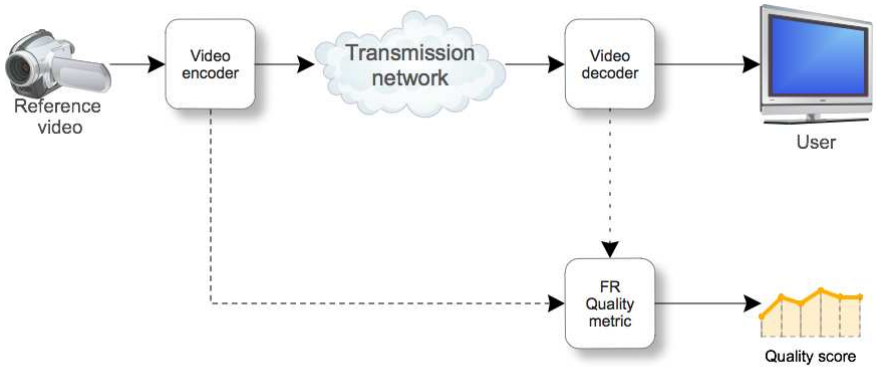
Figure 5: Classification of the video quality metrics according to the input.

metrics require the entire reference video to be available, usually in uncompressed form, which is quite an important restriction on the practical usability of such metrics. Furthermore, full-reference metrics generally impose a precise spatial and temporal alignment of the two videos, so that every pixel in every frame can be matched with its counterpart in the reference clip. Temporal registration in particular is quite a strong restriction and can be very difficult to achieve in practice. Aside from the issue of spatio-temporal alignment, full-reference metrics usually do not respond well to global shifts in brightness, contrast or color, and require a corresponding calibration.

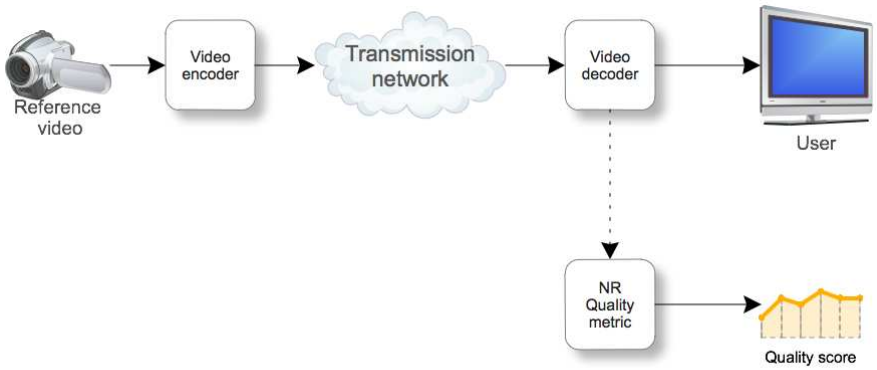
(b) **No-reference:** Also known as "blind" quality assessment, this algorithms are the ideal ones for the rating of transmission, but also the most difficult to find. The problem of making a difference between transmission errors and content of the original videos appears. As can be seen in Figure 6b no-reference (NR) metrics analyze only the video under test, without the need of an explicit reference. This makes them much more flexible than FR metrics, as it can be next to impossible to get access to the reference (e.g. video captured by a camera). They are also completely free from alignment issues. The main difficulty of NR metrics lies in telling apart distortions from content, a distinction humans are usually able to make from experience. NR metrics always have to make assumptions about the video content and/or the distortions of interest. With this comes the risk of confusing actual content with distortions (e.g. a chessboard may be interpreted as block artifacts in the extreme case). The majority of NR metrics are based on estimating blockiness, which is the most prominent artifact of block-DCT based compression methods such as H.26x, MPEG and their derivatives.

(c) **Reduced-reference:** A compromise between FR and NR metrics is taken. As shown in Figure 6c reduced-reference (RR) metrics extract a number of features from the reference video (e.g. the amount of motion or spatial detail), and the comparison with the video under test is then based only on those features. This

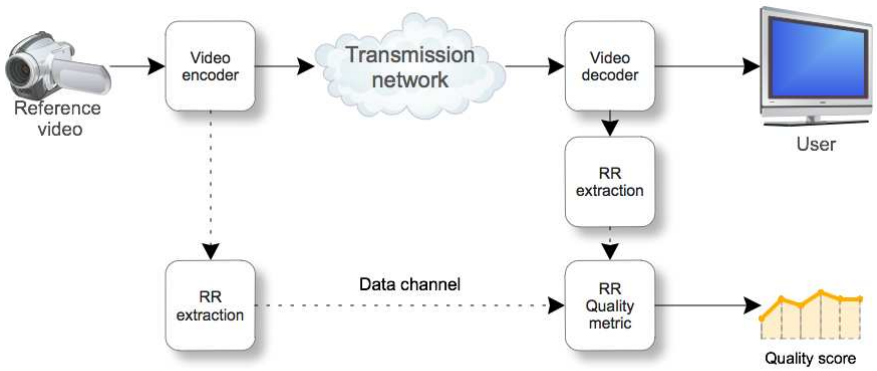
makes it possible to avoid some of the pitfalls of pure no-reference metrics while keeping the amount of reference information manageable. Reduced-reference metrics also have alignment requirements, but they are typically less stringent than for fullreference metrics, as only the extracted features need to be aligned.



(a) Full-reference



(b) No-reference



(c) Reduced-reference

Figure 6: Classification of the video quality metrics according to the availability of the original sequence.

2.3 PSNR

PSNR is a full-reference picture-based metric. It is an engineering term for the ratio between the maximum possible power of a signal, in our case an image, and the power of corrupting noise that affects the fidelity of its representation. Because many signals have a very wide dynamic range, PSNR is usually expressed in terms of the logarithmic decibel scale.

It is most easily defined via the mean squared error (MSE) ,which calculates the error for two monochrome images, where one of the images is considered a noisy approximation of the other.

To understand the MSE and PSNR-algorithm we need first of all to understand how digital images and videos look like. A video sequence is composed by several images one after the other, also known as frames. This frames are divided in a number of pixels depending on the resolution of the video.
In black and white videos each of this pixels has an intensity which varies from 0 to 255 depending on the brightness, 0 for black and 255 for white. The values in the middle are in the grey-scale. In color videos each video has more than one value. By our test-sequences we had 3 values for each, one for the red intensity, one for the green and other for the blue one. This is known as RGB-codification. Since the MSE and PSNR algorithms are thought for black and white images, every RGB pixel value has to be translated to a luminance value. Following the ITU-R Recommendation BT.601 [12] this conversion is calculated as follows:

$$\begin{bmatrix} Y' \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.14713 & -0.28886 & 0.436 \\ 0.615 & -0.51499 & -0.10001 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{1}$$

where Y' is the luminance component and, U and V, are the chrominance comoponents.

The difference between the weights in the formula lies on the fact that the human visual system (HVS) has less receptors for the blue than for the red and for the green.

In statistics, the mean square error or MSE of an estimator is one of many ways to quantify the difference between an estimator and the true value of the quantity being estimated. The MSE measures the average of the square of the error. The error is the amount by which the estimator differs from the quantity to be estimated. The difference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimate.

The MSE of an estimator $\hat{\theta}$ with respect to the estimated parameter θ is defined as:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] \tag{2}$$

In image-processing the MSE of two different frames is calculated as follows:

$$MSE = \frac{1}{n \cdot m} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \left\| I(i,j) - \tilde{I}(i,j) \right\|^2 \tag{3}$$

where $I(i,j)$ denotes the original frame and $\tilde{I}(i,j)$ denotes the distorted frame in pixel position (i,j).

The MSE algortithm is based on the simply basis that if two images are taken into account the difference between them will tell the error and it doesn't matter where the distortion takes place. This might be a problem as will be explained later in the limitations

section. For example, if two images of a football game are taken, where the same distortion is in one of the images on a green part of the field where nothing is happening and in the other image is on the ball hiding it. Since the distortion is exactly the same the MSE algorithm will evaluate the both as the same but since for the viewer the most important part of the image is missing he will not rate both the same.

The PSNR-algorithm is a relative primitive but easy to implement metric for the image quality analysis. It describes the signal-disorder behavior of a image, which means that it gives an idea of how strong the original image was influenced by the noise source. If the result is infinite, it means that there wasn't any distortion, on the other hand, if it is 0 the image is completely distorted, the difference is maximal. A result of 0 would be to expect when for example comparing a black and a white image or frame. Since the video data bases are just error distorted sequences taken from the error-free ones, it is improbable that such a case is given.

PSNR is calculated as follows:

$$PSNR = 10 \cdot \log_{10}\left(\frac{I_{max}^2}{MSE}\right) = 20 \cdot \log_{10}\left(\frac{I_{max}}{\sqrt{MSE}}\right) \tag{4}$$

where Imax is the frame maximum intensity, 255 in our test-sequences.

2.4 SSIM

Since human visual perception is highly adapted for extracting structural information from an image, an alternative metric for video quality assessment will be introduced. It is based on the degradation of structural information and it needs an initial uncompressed or distortion-free video as reference, therefore is is a full-reference video quality metric. It follows a picture-based engineering approach, whis is based on the extraction and analysis of certain features or artifacts in the video. For this purpose a so-called Structural Similarity Index (SSIM) index, has been developed.

Natural images are highly structured: their pixels exhibit strong dependencies, which carry important information about the structure of the persons, animals or objects in the scene. This phenomenom is even more appreciable in spatially proximate pixels. Although most quality measures based on error sensitivity decompose image signals using linear transformations, these do not remove the strong dependencies. To find a more direct way to compare the structues of the reference and the distorted images the SSIM metric was proposed in [13].

Based on the fact that the human visual system is highly adapted to extract structural infomation from a scene, and taking into account that the changes in structural information in an image can provide a good approximation to perceived image distortion, in [14] and [15], a new framework for the design of image quality measures was proposed.

Distortion	MSE	PSNR	SSIM
Sharp	90.1498	28.5812	0.6989
Blur	85.3959	28.8164	0.1868
Motion	85.8508	28.7934	0.2762
JPEG	85.1080	28.8311	0.4740
Salt & pepper	82.3877	28.9722	0.2786

Table 1: Comparison of the same frame with different distortions.

To have a clearlier image of this new way of thinking a comparision with the error sensitivity philosophy (MSE/PSNR) will be done. First of all, the error sesitivity approach

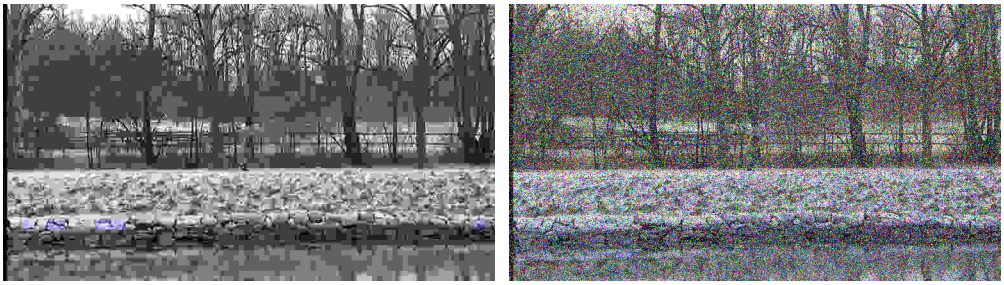
makes an estimation of the quantity of perceived errors to quantify the image degradation, while the structural approachment considers image degradations as pereceived changes in structural information variation.

An example is shown in Figure 7. A reference image is processed with different distortions, each adjusted to have approximately the same PSNR value relative to the reference image. Eventhough they have the same PSNR the images have drastically different perceptual quality. On the one hand, from the error sensitivity point of view, it is difficult to explain why the sharpened or the JPEG compressed images have higher quality in consideration of the fact that its visual difference from the reference image is easily discerned. But it is easily understood with the new philosophy since nearly all the structural information of the reference image is preserved. On the other hand, some structural information from the original image is permanently lost in the blurred and motioned images, and therefore they should be given lower quality scores than the sharped and JPEG compressed images. Regarding the image distorted with salt and pepper impulsive noise, although some of the structures can be discerned it is difficult for the SSIM algorithm to recognize the edges since lots of pixels are lost. The results are reported in Table 1.

Second, the PSNR follows a bottom-up approach, simulating the function of relevant early-stage components in the HVS. The new philosophy is a top-down approach, mim-

(a) Original image (b) Sharped image

(c) Blurred image (d) Motion image



(e) JPEG compressed image (f) Salt & pepper noise image

Figure 7: Comparison of images with different distortions, all with PSNR = 28 dB and a resolution of 768 × 432.

icking the hypothesized functionality of the overall HVS and proposes to evaluate the structural changes between two complex-structured signals directly.

A specific example of a SSIM quality measure from the perspective of image formation has been constructed. The luminance of the surface of an object being observed is the product of the illumination and the reflectance, but the structures of the objects in the scene are independent of the illumination. Consequently, to explore the structural information in an image, the influence of the illumination needs to be separated. The structural information in an image as well as those attributes that represent the structure of objects in the scene are defined without considerating the average luminance and contrast. Instead, since luminance and contrast can vary across a scene, the local luminance and contrast are used for the definition.

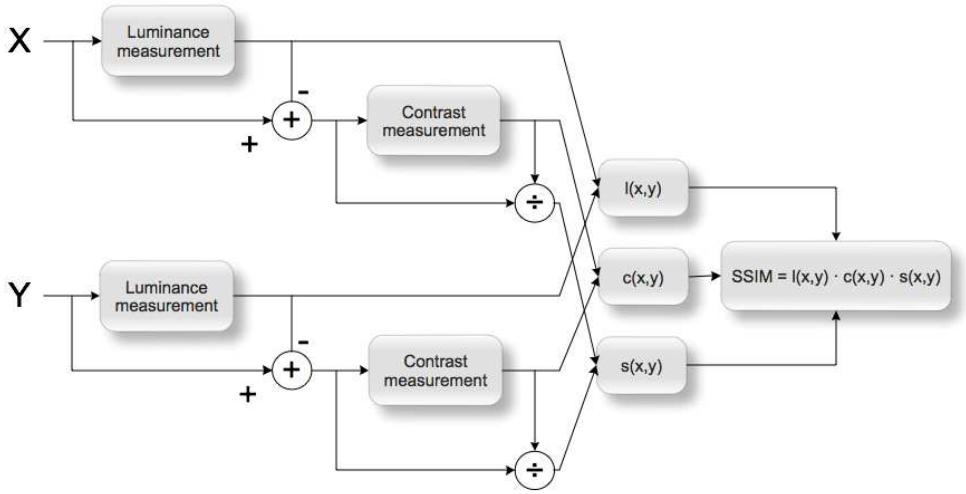


Figure 8: Diagram of the SSIM measurement system.

A system diagram of the metric is shown in in Figure 8. Take x and y, two non-negative image signals, which, previously, have been aligned with each other (e.g., spatial patches extracted from each image). Let's consider x to have perfect quality and y to be distorted. SSIM can serve as a quantitative measurement of the quality of the impaired signal. As seen in the diagram, the system separates the measurement in three parts: luminance, contrast and structure. First of all, the luminance of each image is measured. Assuming discrete signals, this is estimated as the mean intensity:

$$\mu_x = \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} x(i,j) \tag{5}$$

Therefore the luminance comparison $l(x, y)$ function is a function of μ_x and μ_y.

Second, we remove the mean intensity from the signal. In discrete form, the resulting signal x - μ_x corresponds to the projection of vector x onto the hyperplane defined by

$$\sum_{i=1}^{M} \sum_{j=1}^{N} x(i,j) = 0 \tag{6}$$

We use the standard deviation (the square root of variance) as an estimate of the signal contrast. An unbiased estimate in discrete form is given by

$$\sigma_x = \left(\frac{1}{M \cdot N - 1} \sum_{i=1}^{M} \sum_{j=1}^{N} (x(i,j) - \mu_x)^2 \right)^{\frac{1}{2}} \tag{7}$$

The contrast comparison $c(x, y)$ is then the comparison of σ_x and σ_y.

Third, the signal is normalized (divided) by its own standard deviation, so that the two signals being compared have unit standard deviation. The structure comparison $s(x, y)$ is conducted on these normalized signals $(x - \mu_x)/\sigma_x$ and $(y - \mu_y)/\sigma_y$.

Finally, the three components are combined to yield an overall similarity measure

$$S(x, y) = f(l(x, y), c(x, y), s(x, y)) \tag{8}$$

Since the three components (luminance, contrast and structure) a change in luminance and/or contrast does not involve a change in the structures of images. In order to complete the definition of the similarity measure in (8), we need to define the three functions $l(x, y)$, $c(x, y)$, and $s(x, y)$, as well as the combination function $f(\cdot)$. The similarity measure has to satisfy the following conditions:

1. Symmetry: $S(x, y) = S(y, x)$.
2. Boundedness: $S(x, y) \leq 1$.
3. Unique maximum: $S(x, y) = 1$ if and only if $x = y$ (in discrete representations, $x(i, j) = y(i, j)$ for all i =1, 2, ..., M and j =1, 2, ..., N).

Luminance comparison is defined as

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{9}$$

where the constant C_1 is included to avoid instability when $\mu_x^2 + \mu_y^2$ is very close to zero. Specifically

$$C_1 = (K_1 L)^2 \tag{10}$$

where L is the dynamic range of the pixel values (255 for 8-bit grayscale images), and $K_1 \ll 1$ is a constant. Similar considerations also apply to contrast comparison and structure comparison described later. Equation (9) is easily seen to obey the three properties listed above. Equation (9) is also qualitatively consistent with Weber's law, which has been widely used to model light adaptation (also called luminance masking) in the HVS. According to Weber's law, the magnitude of a just-noticeable luminance change Δ is approximately proportional to the background luminance I for a wide range of luminance values. In other words, the HVS is sensitive to the relative luminance change, and not the absolute luminance change. Let R represent the size of luminance change relative to background luminance, we rewrite the luminance of the distorted signal as $\mu_y = (1 + R)\mu_x$. Substituting this into (9) gives

$$l(x, y) = \frac{2(1 + R)}{1 + (1 + R)^2 + \frac{C_1}{\mu_x^2}} \tag{11}$$

If we assume C_1 is small enough (relative to μ_x^2) to be ignored, then $l(x, y)$ is a function only of R, qualitatively consistent with Weber's law.

The contrast comparison function takes a similar form

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{12}$$

where $C_2 = (K_2L)^2$, and $K_2 \ll 1$. This definition again satisfies the three properties listed above. In the case of same amount of contrast change $\Delta\sigma = \sigma_y - \sigma_x$, this measure is less sensitive to the case of high base contrast σ_x than low base contrast, which is consistent with the contrast-masking feature of the HVS.

Structure comparison is doner after luminance subtraction and variance normalization of both signals. The two unit vectors $(x - \mu_x)/\sigma_x$ and $(y - \mu_y)/\sigma_y$, each lying in the hyperplane defined by (6), are associated with the structure of the two images. The correlation between these is a simple and effective measure to quantify the structural similarity. Since the correlation between $(x - \mu_x)/\sigma_x$ and $(y - \mu_y)/\sigma_y$ is equivalent to the correlation coefficient between x and y, the structure comparison function is defined as follows:

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{13}$$

As in the luminance and contrast measures, a small constant in both denominator and numerator has been introduced. In discrete form, σ_{xy} can be estimated as

$$\sigma_{xy} = \frac{1}{M \cdot N - 1} \sum_{i=1}^{M} \sum_{j=1}^{N} (x(i,j) - \mu_x)(y(i,j) - \mu_y) \tag{14}$$

Geometrically, the correlation coefficient corresponds to the cosine of the angle between the vectors $(x - \mu_x)$ and $(y - \mu_y)$. Note also that $s(x,y)$ can take on negative i's.

Finally, the three comparisons of (9), (12) and (13) are combined

$$SSIM(x,y) = [l(x,y)]^\alpha \cdot [c(x,y)]^\beta \cdot [s(x,y)]^\gamma \tag{15}$$

where $\alpha > 0$, $\beta > 0$ and $\gamma > 0$ are parameters used to adjust the relative importance of the three components. It is easy to verify that this definition satisfies the three conditions given above. In order to simplify the expression $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$ have been set. This results in a specific form of the SSIM index

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{16}$$

As can be seen in [16] and [17] , the "universal quality index" (UQI) corresponds to the special case that $C_1 = C_2 = 0$, which produces unstable results when either $(\mu_x^2 + \mu_y^2)$ or $(\sigma_x^2 + \sigma_y^2)$ is very close to zero.

For image quality assessment, it is useful to apply the SSIM index locally rather than globally. First, image statistical features are usually highly spatially nonstationary. Second, image distortions, which may or may not depend on the local image statistics, may also be space-variant. Third, at typical viewing distances, only a local area in the image can be perceived with high resolution by the human observer at one time instance (because of the foveation feature of the HVS [18], [19]). And finally, localized quality measurement can provide a spatially varying quality map of the image, which delivers more information about the quality degradation of the image and may be useful in some applications.

In [16] and [17], the local statistics μ_x, σ_x and σ_{xy} are computed within a local 8 $\times$ 8 square window, which moves pixel-by-pixel over the entire image. At each step, the local statistics and SSIM index are calculated within the local window. One problem with this method is that the resulting SSIM index map often exhibits undesirable

"blocking" artifacts. Therefore a 11 × 11 circular-symmetric Gaussian weighting function $w = \{w(i,j)|i,j = 1, 2, \ldots, M, N\}$ is used, with standard deviation of 1.5 samples, normalized to unit sum ($\sum_{i=1}^{M}\sum_{j=1}^{N} w(i,j)$). The estimates of local statistics μ_x, σ_x and σ_{xy} are then modified accordingly as

$$\mu_x = \sum_{i=1}^{M}\sum_{j=1}^{N} w(i,j)x(i,j) \tag{17}$$

$$\sigma_x = \left(\sum_{i=1}^{M}\sum_{j=1}^{N} w(i,j)(x(i,j) - \mu_x)^2\right)^{\frac{1}{2}} \tag{18}$$

$$\sigma_{xy} = \sum_{i=1}^{M}\sum_{j=1}^{N} w(i,j)(x(i,j) - \mu_x)(y(i,j) - \mu_y) \tag{19}$$

With such a windowing approach, the quality maps exhibit a locally isotropic property. Throughout this thesis, the SSIM measure uses the following parameter settings: K_1 =0.01; K_2 =0.03. These values are somewhat arbitrary, but it has been found out that, the performance of the SSIM index algorithm is fairly insensitive to variations of these values.

In practice, one usually requires a single overall quality measure of the entire image. A mean SSIM (MSSIM) index to evaluate the overall image quality will be used

$$MSSIM(X,Y) = \frac{1}{L}\sum_{k=1}^{L} SSIM(x_k, y_k) \tag{20}$$

where X and Y are the reference and the distorted frames, respectively; x_k and y_k are the frame contents at the kth local window; and L is the number of local windows of the image. Depending on the application, it is also possible to compute a weighted average of the different samples in the SSIM index map. For example, region-of-interest image processing systems may give different weights to different segmented regions in the image. As another example, it has been observed that different image textures attract human fixations with varying degrees (e.g., [20], [21]). A smoothly varying foveated weighting model (e.g., [19]) can be employed to define the weights. In this thesis, however, use uniform weighting will be used. A MATLAB implementation of the SSIM index algorithm is available online at [22].

2.5 VQM

Video quality metric general model (VQM) is a reduced reference metric that combines different parameters using linear models to produce estimates of video quality that closely approximate subjective test results [24]. The VQM metric was tested in the Video Quality Experts Group (VQEG) Phase II Full Reference Television (FR-TV) tests [23]. In these tests VQM was the only model that performed statistically better than the other in both the 525-line and 625-line tests. As a result VQM was standart sized by ANSI in July 2003 (ANSE T1.801.03-2003), and has been included in Draft Recommendations from ITU-T Study Group 9 and ITU-R Working Party 6Q.

Pinson and Wolf's video quality metric [24] divides sequences into spatio-temporal blocks, and a number of features measuring the amount and orientation of activity in each of these blocks are computed from the spatial luminance gradient. The features extracted from test and reference videos are then compared using a process similar to masking.

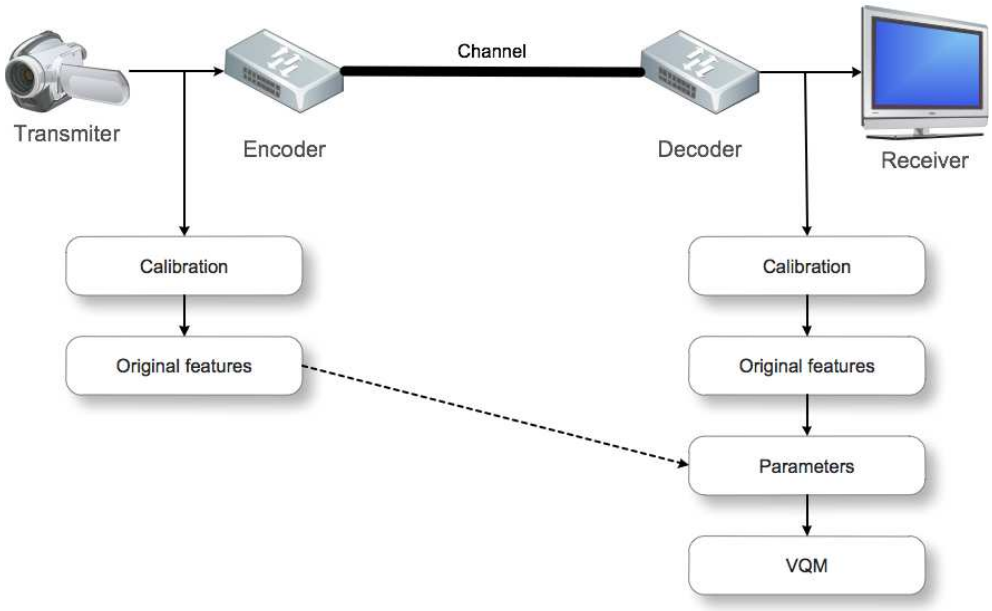


Figure 9: Block diagram of the VQM measurement system.

The VQM utilizes reduced-reference picture-based, following an engineering approach, parameters that are extracted from optimally-sized spatial-temporal (S-T) regions of the video sequences. These features are extracted from the source and distorted video streams. Therefore a calibration for comparing both videos in the same environment is needed. VQM and its associated calibration techniques follow a complete automated process that can be seen in Figure 9. The calibration of the original and distorted video streams consists of four parts: spatial alignment, valid region estimation, gain and level offset calculation and temporal alignment. Then the perception-based features are extracted and the video quality parameters are computed. Finally these parameters are combined to obtain the estimate of video quality.

In the spatial alignment of the calibration process the horizontal and vertical spatial shift of the distorted video relative to the original video is determined. In this case the accuracy of the spatial alignment algorithm is to the nearest 0.5 pixel for horizontal shifts and to the nearest line for vertical shifts. Once the spatial alignment has been calculated, the spatial shift is removed from the distorted video stream. In case of interlaced video a reframing process might be inclueded. Since spatial alignment must be determined before the distorted valid region, gain and level offset and temporal alignment, and each of those quantities must be computed by comparing original and distorted video contents that have been spatiall registered a "chicken or egg" measurement problem appears. The interdependence of these processes makes that an exhaustive search over all variables would require an enormous number of computations. The solution presented in the paper previously mentioned [24] is to perform an iterative search to find the closest matching original frame for each distorted frame. The spatial alignment algorithm described above requires a relatively high data channel bandwhidth, due to the pixel-by-pixel comparison of original and distorted frames. In case of a monitoring application this would involve a more accurate design of it. Fortunately, each piece of video transmission equipment will nomarlly have one constant spatial alignment.

According to ITU-R Recommendation BT.601 [12] sampled videos may have a border

of pixels and lines that do not contain a valid picture. To prevent non-picture areas from influencing the VQM measurements, these areas are excluded from the VQM measurement. The processed valid region (PVR) is calculated for each scene separately and then the invalid pixels are discarded from the original and distorted video sequences. An automated algorithm has been developed. It estimates the valid region of the distorted video stream so that subsequent computations do not consider corrupted lines at the top and bottom of the Rec. 601 frame, black border pixels, or transitional effects where the black border meets the picture area.

Once original and distorted frames are spatially and temporally registered, the gain and level offset calibration can be performed on either fields or frames as appropriate. The method used makes the assumption that the Rec. 601 Y, U and V signals each have an independent gain and level offset and will not properly calibrate video systems that introduce a phase rotation of the chrominance information. Although gain and level offsets are calculated for the U and V channels, these correction factors will not be applied. VQM will only use the Y channel gain and level offset correction factors.

Video delay can depend upon dynamic attributes of the original scene and video system Therefore estimates of video delay are required to temporally align the original and processed video streams before making quality measurements. Some video transimission systems may provide time synchronization information, but, in general, time synchronization between the original and processed video streams must be measured. The technique used for VQM is "frame-based" in that it works by correlating lower resolution images, sub-sampled in space and extracted from the original and distorted video streams. Then the delay of each frame or field is estimated and finally these individual estimates are combined to estimate the average delay of the video sequence.

The next step in the VQM process is to extract the features of the original and distorted videos. By comparing the features extracted from the calibrated distorted video with features extracted from the original video, quality parameters can be computed.

On the one hand all of the quality features used by VQM quantify the perceptual aspect of a video stream by performing the following steps. First, a perceptual filter is applied to the video stream to enhance some property of perceived video quality. Second, features are extracted from spatial-temporal (S-T) sub-regions using a mathematical function. S-T region sizes are described by the number of pixels horizontally, the number of frame lines vertically, and the time duration of the region and the most used functions are the mean and the standard deviation. And third, by clipping some features values a perceptibility thresholsd is applied to the extracted feature stream.

$$f_{clip} = max(f, threshold) \tag{21}$$

This clipping serves to reduce sensitivity to imperceptible impairments.

On the other hand quality parameters compare original and distorted features to obtain an overall measure of video distortion by following this steps. One, the distorted feature value for each S-T region is compared to the corresponding original feature value using comparison functions that emulate the perception of impairments. This functions can be one of the following:

1. Euclidean distance:
$$p = \sqrt{(f_o - f_p)^2 + (f_{o2} - f_{p2})^2} \tag{22}$$

2. Ratio comparison:
$$p = (f_p - f_o)/f_o \tag{23}$$

3. Log comparison:

$$p = log_{10}\left(\frac{f_p}{f_o}\right) \tag{24}$$

Where f_o and f_{o2} are original feature values; f_p and f_{p2} are the corresponding distorted feature values. The ratio and log comparison functions produce a mixture of positive (gains) and negative (losses) values. Greater measurement accuracy can be obtained by examining losses and gains separately. The reason is that humans react more negatively to additive impairments than subtractive impairments and hence losses and gains must be given different weights in the quality estimator. Therefore the ratio and log funcions are always followed by a either a loss funcion (replace positive values with zero) or a gain function (replace negative values with zero). Two, perception-based error-pooling functions are applied across space (spatial collapsing) and time (temporal collapsing). Three, the final space-time collapsed parameter values may also be clipped to account for nonlinearities and to better match the parameter's sensitivity to impairments with human perception of those impairments. In this case the clipping function looks like this:

$$p' = \begin{cases} 0 & \text{if } p \leq t \\ p - t & \text{otherwise} \end{cases} \tag{25}$$

VQM is composed by seven independent parameters. Four are based on features extracted from spatial gradients of the Y componnent, two on features extracted from the vector formed by the two U and V chrominance components, and one on the product of features that measure contrast and motion. The parameters are computed as described below:

- Parameter "**si_loss**"

  Parameter si_loss detects a decrease or loss of spatial information. It uses a 13 pixel spatial information filter (SI13), which was specifically developed to measure perceptually significant edge impairments [25]. SI13 utilizes 13 pixel by 13 pixel horizontal and vertical filter masks. These two filter masks are created by the replication of the following vector: [-.0052625, -.0173446, -.0427401, -.0768961, -.0957739, -.0696751, 0, .0696751, .0957739, .0768961, .0427401, .0173446, .0052625]

  The filters are applied separately to the liminance frame resulting two filtered images (I_H and I_V), which are combined into a single image (I_{SI13}) using Euclidean distance.

- Parameter "**hv_loss**"

  Parameter hv_loss detects a shift of edges from horizontal and vertical orientation to diagonal orientation. It uses the horizontally and vertically filtered H and V images from the SI13 filter. Two new perceptually filtered images are created: one contains horizontal and vertical edges (HV) and the other contains diagonal edges (HV_{BAR}, or complement of HV).

- Parameter "**hv_gain**"

  This parameter detects a shift of edges from diagonal to horizontal and vertical.

- Parameter "**chroma_spread**"

  This parameter detects changes in the spread of the distribution of two-dimensional color samples.

- Parameter "**si_gain**"

  This is the only quality improvement parameter in the model. The si_gain parameter measures improvements to quality that result from edge sharpening or enhancements.

- Parameter "**ct_ati_gain**"

  This metric is computed as the product of a contrast feature, measuring the amount of spatial detail, and a temporal information feature, measuring the amount of motion present in the S-T region. Impairments will be more visible in S-T regions that have a low product than in S-T regions that have a high product. ct_ati_gain identifies moving-edge impairments that are nearly always present.

- Parameter "**chroma_extreme**"

  This metric detects severe localized color impairments, such as those produced by digital transmission errors.

VQM consists of a linear combination of the parameters described above. It outputs values that range from zero (no perceived impairment) to approximately one (maximum perceived impairment). Therefore VQM is defined as:

$$\begin{aligned} VQM = &- 0.2097 \cdot si\_loss + 0.5969 \cdot hv\_loss + 0.2483 \cdot hv\_gain + 0.0192 \cdot chroma\_spread \\ &- 2.3416 \cdot si\_gain + 0.0431 \cdot ct\_ati\_gain + 0.0076 \cdot chroma\_extreme \end{aligned} \tag{26}$$

Note that si_loss is always less than or equal to zero, so si_loss can only increase VQM. Since all the other parameters are greater than or equal to zero, si_gain is the only parameter that can decrease VQM.

VQM is clipped at a lower threshold of 0.0. This prevents si_gain values from producing a quality rating that is better than the original (i.e., a negative VQM). Finally, a crushing function that allows a maximum 50% overshoot is applied to VQM values over 1.0. The purpose of the crushing function is to limit VQM values.

If VQM $\leq$ 1.0, then VQM = (1 + c) * VQM / (c + VQM), where c= o.5

VQM computed like above will have values between zero and one. It might be occasionally greater than one in video scenes that are extremely distorted.

2.6 VQuad

VQuad is a full-reference picture-based metric developed by SwissQual in 2008. VQuad08 improves several and extends detectors for individual degradation. It is more robust to the latest coding technologies and less content dependent. It also provides a set of additional results giving more details about the type of distortions that came up during the analysis of the video sequences. Therefore an easier interpretation and localization of potential problems in quality estimation is available.

VQuad is able to identify the following perceptual degradations [26], which later will be used to predict the MOS values:

- **Blockiness**

  Nowadays almost all video encoders use a block based transform. In this process the images are divided in small squares called blocks causing the so called blockiness effect. A lossy encoding of these blocks may cause that a resulting block structure

can be seen in the decoded video sequence. The most typical block sizes are 8×8 and 16×16 pixels.

The block based transform is a DCT transformation, which encodes the luminance and chrominance information separately, even with different block sizes. At the end, only the most significant coefficients of the transformation are retained. In case of a strong compression very few coefficients will be used, in extreme cases just one of them will be retained. Normally the one representing a uniform color or luminance of the whole block. This means that the whole block will look the same leading to less or no spatial detail inside it but to visible transitions between blocks. Since no transition details are available the border area to the neighboring blocks becomes more visible.

The blockiness value is estimated by measuring the luminance differences at block borders. It is related to the amount of spatial detail, since a block border has a stronger visibility in the absence of spatial details. Blocks might have different sizes, however the borders will always be horizontally or vertically oriented and therefore form a right angle. In very bright or dark areas the degradation between borders is less visible eventhough it is clearly measureable.

The main reason for the blockiness effect is the strong compression during the encoding processes. Furthermore, packet loss during transmission could increase blockiness.

- **Tilling** Tilling is the effect of visible tile-like artifacts in the video frames. It can happen due to either the encoding process or the transmission.

  The tilling value is based on the distortions at block borders caused by transmission errors. This type of error can be handled by the receiving decoder following different strategies. The simplest one is just to display the erroneus data, which leads to very strange effects. A more developed strategy is to freeze the last succesfully updated video frame up to the next key-frame providing a complete image at once. Another way to deal with this kind of errors is to replace the erroneus transmitted parts of the video frame by the same area of the previous video frame, this is known as slicing. Advanced strategies predict the missing data by the neighboring blocks. Since no concealment strategy is perfect, the residual error will be propagated by the following differential frames up to the next key-frame.

  The smallest entity in the transmission of video is a macro-block, therefore an error in the process often causes a kind of macro-block driven visible structure. As said above the border-lines between blocks are horizontally or vertically oriented. Because of that the tilling detector is designed to recognize the erroneus areas by checking in-coherent vertical and horizontal edges. To avoid false detections and thus lowered scores a threshold is applied.

  In previous versions of VQuad the visible macro-block borders caused by spatial compression were counted as tilling. With the latest VQuad08 version the blockiness value includes these distortions caused by the spatial compression. However, suddenly appearing macro-block structures by a highly compressed key-frame or temporarily increased spatial compression might be considered as tiling.

  The main reason for the tiling effect is packet loss during transmission. Furthermore, strong compression of encoding might increase tiling.

- **Blurring**

In VQuad08 blurring is measured indirecly by the measuring of sharpness, which measures the luminance offset at edge borders in the video frames and relates this to the local contrast at the edge location. Sharpness tries to avoid edges which form block borders as a result of blocking or tilling.

The blurring value is the decrease of the sharpness of the reference video frame to the transmitted one. Sharpness is strongly content dependent, therefore the sharpness is measured at the position of the sharpest edges in the video frame.

The main reason for blurring is the use of de-blocking filters of the video decoder.

- **Jerkiness**

  Jerckiness is the result of bad representation of moving objects in a video sequence. Jerckiness is a perceptual degradation, which measures the loss of information due to freezing or low frame rates. Therefore freezing and the 'Dominating Frame Rate' along with this anticipated loss of information forms the jerkiness.

  In a freezing period jerckiness considers the loss of information during the period. This loss is estimated by the inter-frame difference at the end of the period.

  When no freezing periods are present, jerckiness is mainly related to the 'Dominating Frame Rate'. Since jerkiness and frame rate are highly negatively correlated, only for sequences with slow motion a low frame rate does not imply jerckiness. The reason is that the jerckiness measure takes into account the amount of motion in the video.

  Large jerkiness values are a result of the reduction to a low frame rate at the encoder, or to transmission delays and strong packet loss, respectively, during transmission.

- **Perceptual difference**

  Since VQuad08 is a full-reference method, it has access to the reference video sequences. This permits a detailed comparison of the reference and transmitted video sequences. For the calculation of the perceptual difference the frames of the reference and transmitted sequences have to be time aligned. This means that for each frame of the reference sequence an identical frame of the transmitted sequence exists. It might happen that due to distortions no frame from the reference sequence can be assigned. Such cases are known as unmatched frames.

  Once the frames are aligned, the perceptual difference is calculated between the corresponding frames. The resulting value is a key parameter for the predicted MOS value. This value is obtained by calculating the interframe difference between matched frames. Emphasis is putted on large edges and adaptation effects to luminance and local contrast.

The above described perceptual degradation measures are used to estimate a MOS prediction. As those are non-additive and therefore non-linear, the most important degradation will determine the objective quality, while the less important distortions will have a smaller weight in the predicted MOS.

$$MOS_{predicted} = f_{temporal}(video) \cdot f_{spatial\_FR}(video) \tag{27}$$

where $f_{temporal}$ is a function of temporal degradations and $f_{spatial\_FR}$ is a function of spatial degradations dominated by the perceptual difference measure.

2.7 MCE

Another interesting approach is the one suggested in [27], from now on the MCE metric. It is a no-reference hybrid based metric, which makes the tool easy to incorporate into existing systems. Therefore this no-reference metric uses two types of information: the reconstructed pixels and the bitstream.

MCE video quality metric uses a hybrid of signal pixel-based and bitstream-based video quality metric that can estimate video quality degradation caused by packet loss. Three steps are followed by the algorithm to calculate the video quality based on the error-concealment effeciveness. First of all the macroblocks containing errors in each frame are detected using the bitstream information. Second, the effectiveness of error concealment is evaluated using bitstream and decoded frame information. Third, the video decoder of the system outputs the estimated mean square error (MSE) values.

For the first two steps two kinds of information are required. Firstly, the degree of motion in the scene, which will be given by the bit-stream information. Two, the luminance discontinuity at boundaries between correctly and error-concealed areas (in both horizontal and vertical directions), which will be extracted from the pixel information. The error-concealment effectiveness will depend on both characteristics. The more degree of motion or luminance discontinuity, the less the error-concealment effectiveness.

- Error-concealment effectiveness using motion information:

  The motion information of the scene will be extracted from the motion vectors in an input bitstream. The degree of motion is represented as a sum of absolute motion vector values. If the sum is larger than the pre-determined threshold, error concealment is considered to be ineffective.

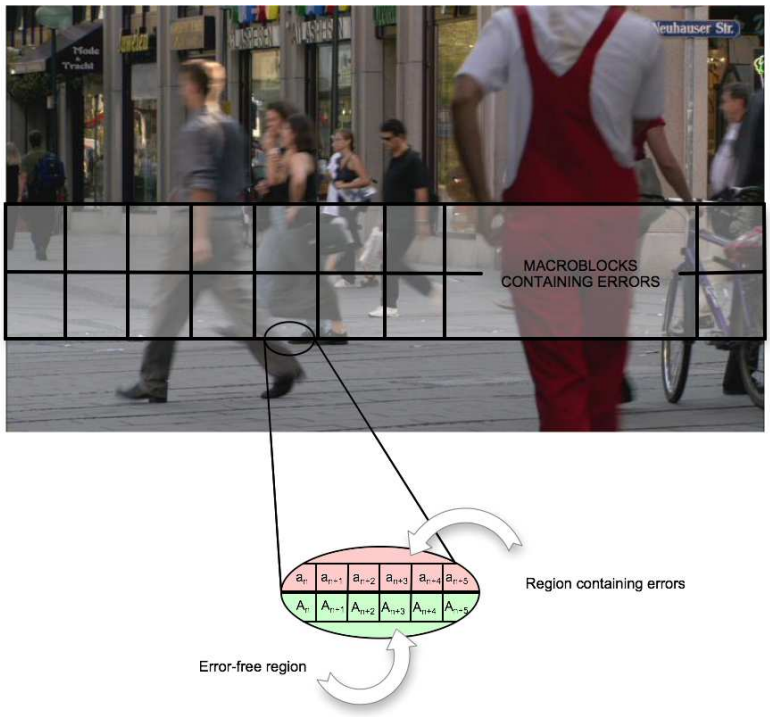


Figure 10: Pixels along a boundary of an error region for MCE calculation.

$$|MV_x| + |MV_y| > Th_{mv} \tag{28}$$

Where (MV_x, MV_y) is the motion vector for a macroblock in the same position as that of the impairment-macroblock in the previous P-frame. Th_{mv} is the threshold. As discussed in [27] Th_{mv} =2. Any impairment-macroblock for which equation (28) is satisfied or of which a motion vector is not available is considered one for which error-concealment has been ineffective.

- Error-concealment effectiveness using luminance information at error region boundaries:

  Luminance discontinuity is calculated as:

$$D = \frac{1}{N} \sum_{n=0}^{N-1} |a_n - A_n| > Th_L \tag{29}$$

  Where as illustrated in Figure 10, a_n is the luminance value in the error region along the boundary, and A_n is the luminance value in the correctly decoded region. N is the total number of pixels along the boundary of a macroblock. And Th_L is the threshold value. In this case $Th_L = 7$, as in [27]. If D is greater than the threshold Th_L the error concealment will be considered ineffective. This comparison is applied to every boundary between the error-regions and correctly decoded regions.

The amount of video quality degradation is estimated on the basis of the total number of macroblocks for which the error concealment has been found to be ineffective. Therefore an estimated mean square error, MSE_{est}, can be calculated on the basis of error-concealment-ineffective macroblocks, x:

$$MSE_{est} = 0.0023x^2 - 0.0092x + 84.686 \tag{30}$$

2.8 SSIM+

Based on the same principle as SSIM, that human visual perception is highly adapted for extracting structural information from an image, a new metric has been developed, called SSIM+. Since this new metric follows the SSIM principles it needs the original undistorted video as reference, therefore it is a full-reference metric.

As detailed in section 2.4 section, natural images are highly structured. The proposed method tries to utilize that by enhancing the way of calculating the structure coefficient of the SSIM formula.

Since edge detection helps to detect the presence and locations of intensity transitions, it drastically reduces the amount of data needed by the metrics to make an estimation of the video quality. It also provides important information about the shapes of objects and it is easy to integrate into a large number of object recognition algorithms, like the one under discussion.

The calculation of the structure coefficient is based on the Laplacian of a Gaussian. This method finds edges by looking for zero crossings after reducing the sensitivity to noise by filtering I with a Laplacian of Gaussian filter, which also helps on the edge/structure detection. Edge localization is another problem encountered in edge detection. The addition of noise to an image can cause the position of the detected edge to be shifted from its true location. The ability of an edge-detector to locate in noisy data an edge that is as close as possible to its true position in the image is an important factor in determining its performance. Another difficulty in any edge detection system arises from the fact that the sharp intensity transitions which indicate an edge are sharp because of their

high-frequency components. As a result, any linear filtering or smoothing performed on these edges to suppress noise will also blur the significant transitions. However, some form of smoothing is necessary since edge detection depends on differentiating the image function and this amplifies all high-frequency components of the signal, including those of the noise. Low-pass filters are the most widely used smoothing filters. The amount of smoothing applied depends on the size or scale of the smoothing operator. In general, for a small scale, the detector extracts fine details of intensity changes from the image, but tends to be more sensitive to noise. A larger scale extracts coarse details of intensity changes, but some of the detected edges tend to have a large localization error. Selecting a single scale of smoothing which is optimal for all edges in an image is very difficult. One filter size may not be good enough to remove noise while keeping good localization.

The most widely used smoothing filters are Gaussian filters [28]. Such filters have been shown to play an important role in edge detection in the human visual system, and to be extremely useful as detectors for edge and line detection.

By variational methods, Canny derives an optimal edge detection operator which turns out to be well approximated by the first derivative of a Gaussian function. It has also been proved that when one-dimensional (1-D) signals are smoothed with a Gaussian filter, the scale space representation of their second derivatives shows that existing zero-crossings disappear when moving from a fine-to-coarse scale, but new ones are never created.

It is also proved that for a wide category of signals, the Gaussian function is the only filter that has this property. This unique property makes it possible to track zero-crossings over a range of scales, and also gives the ability to recover the entire signal at sufficiently small scales.

Extending this work to two-dimensional (2-D) signals and proved that with the Laplacian, the Gaussian function is the only filter in a wide category that does not create zero-crossings as the scale increases. It has also been showed that for nonlinear directional derivatives along the gradient, there is no filter that does not create zero-crossings as the scale increases. The 2-D Gaussian filter is also the only rotationally symmetric filter that is separable in Cartesian coordinates. Separability is important for computational efficiency when implementing the smoothing operation by convolutions in the spatial domain.

Since the Laplace operator may detect edges as well as noise (isolated, out-of-range), it may be desirable to smooth the image first by convolution with a Gaussian kernel of width σ.

$$G_\sigma(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{x^2+y^2}{2\sigma^2}] \tag{31}$$

to suppress the noise before using Laplace for edge detection:

$$\Delta[G_\sigma(x,y) * f(x,y)] = \Delta[G_\sigma(x,y)] * f(x,y) = LoG * f(x,y) \tag{32}$$

The first equal sign is due to the fact that

$$\frac{d}{dt}[h(t) * f(t)] = \frac{d}{dt}\int f(\tau)h(t-\tau)\,\mathrm{d}\tau = \int f(\tau)\frac{d}{dt}h(t-\tau)\,\mathrm{d}\tau = f(t) * \frac{d}{dt}h(t) \tag{33}$$

So we can obtain the Laplacian of Gaussian $\Delta G_\sigma(x,y)$ first and then convolve it with the input image. To do so, first consider

$$\frac{\delta}{\delta x}G_\sigma(x,y) = \frac{\delta}{\delta x}\exp^{-(x^2+y^2)/2\sigma^2} = -\frac{x}{\sigma^2}\exp^{-(x^2+y^2)/2\sigma^2} \tag{34}$$

and

$$\frac{\delta^2}{\delta^2 x} G_\sigma(x,y) = \frac{x^2}{\sigma^4} \exp^{-(x^2+y^2)/2\sigma^2} - \frac{1}{\sigma^2} \exp^{-(x^2+y^2)/2\sigma^2} = \frac{x^2-\sigma^2}{\sigma^4} \exp^{-(x^2+y^2)/2\sigma^2} \quad (35)$$

Note that for simplicity we omitted the normalizing coefficient $1/\sqrt{2\pi\sigma^2}$. Similarly we can get

$$\frac{\delta^2}{\delta^2 x} G_\sigma(x,y) = \frac{y^2-\sigma^2}{\sigma^4} \exp^{-(x^2+y^2)/2\sigma^2} \quad (36)$$

Now we have LoG as an operator or convolution kernel defined as

$$LoG = \Delta G_\sigma(x,y) = \frac{\delta^2}{\delta^2 x} G_\sigma(x,y) + \frac{\delta^2}{\delta^2 y} G_\sigma(x,y) = \frac{x^2+y^2-\sigma^2}{\sigma^4} \exp^{-(x^2+y^2)/2\sigma^2} \quad (37)$$

The edges in the image can be obtained by these steps:

- Applying LoG to the image.
- Detection of zero-crossings in the image.
- Threshold the zero-crossings to keep only those strong ones (large difference between the positive maximum and the negative minimum).

The last step is needed to suppress the weak zero-crossings most likely caused by noise.

This procedure is applied to both the original and distorted frames. Once the edges/stuctural maps of both images the structure coefficient is calculated:

$$s(x_{LoG}, y_{LoG}) = \frac{\sigma_{xLoGyLoG} + C_3}{\sigma_{xLoG}\sigma_{yLoG} + C_3} \quad (38)$$

Where σ_{xLoG} and σ_{yLoG} are:

$$\sigma_{xLoG,yLoG} = \left(\sum_{i=1}^{M} \sum_{j=1}^{N} w(i,j)(x_{LoG}, y_{LoG}(i,j) - \mu_{xLoG,yLoG})^2 \right)^{\frac{1}{2}} \quad (39)$$

and

$$\sigma_{xy_{LoG}} = \sum_{i=1}^{M} \sum_{j=1}^{N} w(i,j)(x_{LoG}(i,j) - \mu_{xLoG})(y_{LoG}(i,j) - \mu_{yLoG}) \quad (40)$$

The result of applying this procedure to one random frame is shown in Figure 11, where the edges are represented with white lines.

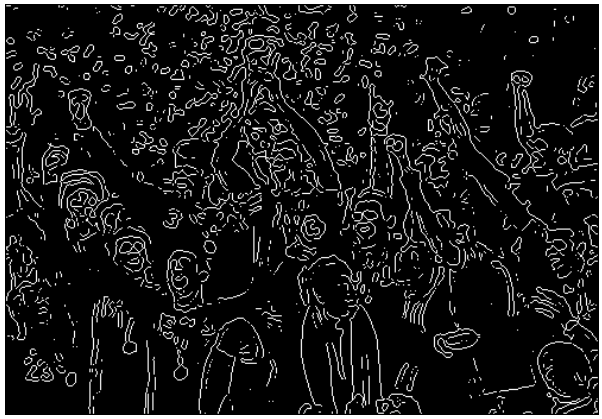
Now that is known how to calculate the structural index $s(x,y)$ we will be able to introduce it in the SSIM formula:

$$SSIM(x,y) = [l(x,y)]^\alpha \cdot [c(x,y)]^\beta \cdot [s(x,y)]^\gamma \quad (41)$$

where $\alpha > 0$, $\beta > 0$ and $\gamma > 0$ are parameters used to adjust the relative importance of the three components. In order to simplify the expression $\alpha = \beta = \gamma = 1$ have been set. A block diagram of how SSIM+ is caculated is shown in Figure 12.

The luminance and the contrast indexs remain the same as in SSIM:

(a) Original frame



(b) Frame after application of LoG

Figure 11: Application of LoG method to a random frame.

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{42}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{43}$$

where the constant C_1 and C_2 are included to avoid instability when $(\mu_x^2 + \mu_y^2)$ or $(\sigma_x^2 + \sigma_y^2)$ are very close to zero.

As was done in SSIM and following the indications of [13] the local stadistics $\mu_{x,y}$, $\sigma_{x,y}$ and σ_{xy} were computed within a local window, which moves pixel-by-pixel through the images. Every time the stadistics are newly calculted using the windowed values. The window is, once again as in SSIM, a 11×11 circular-symmetric Gaussian weighting function $w = \{w(i,j)|i,j = 1, 2, \ldots, M, N\}$ with standard deviation of 1.5 samples, normalized to unit sum $(\sum_{i=1}^{M}\sum_{j=1}^{N} w(i,j))$.

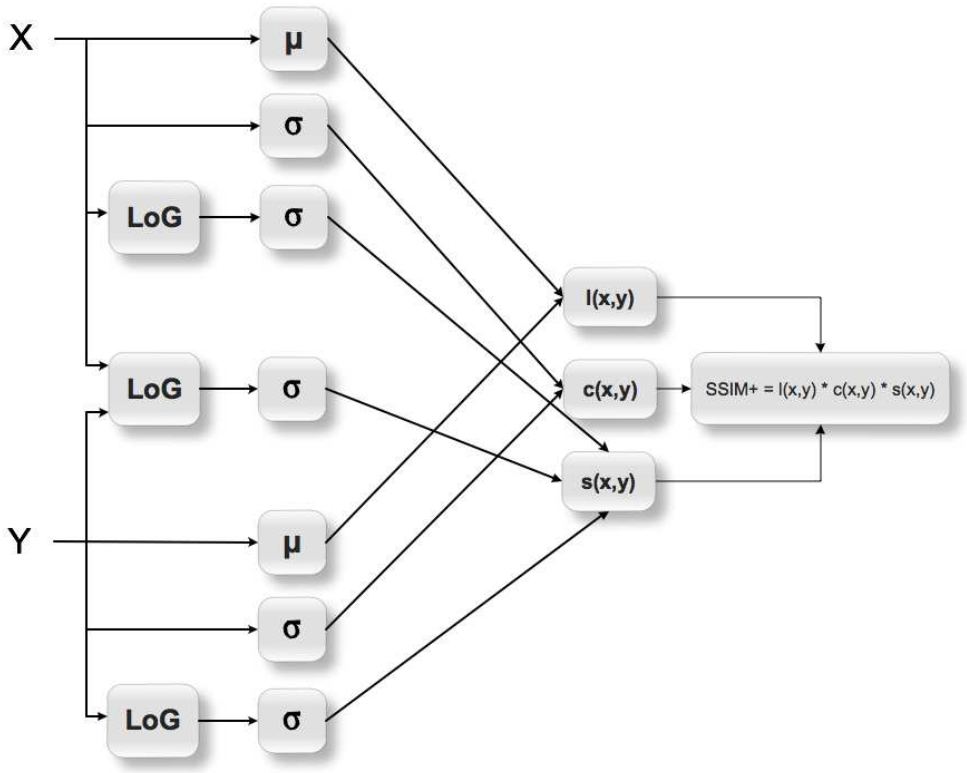


Figure 12: Diagram of the SSIM+ measurement system.

$$\mu_{x,y} = \sum_{i=1}^{M}\sum_{j=1}^{N} w(i,j)x(i,j), y(i,j) \tag{44}$$

$$\sigma_{x,y} = \left(\sum_{i=1}^{M}\sum_{j=1}^{N} w(i,j)(x(i,j), y(i,j) - \mu_{x,y})^2\right)^{\frac{1}{2}} \tag{45}$$

$$\sigma_{xy} = \sum_{i=1}^{M}\sum_{j=1}^{N} w(i,j)(x(i,j) - \mu_x)(y(i,j) - \mu_y) \tag{46}$$

As explained above the stuctural index $s(x, y)$ is newly calculated as:

$$s(x,y) = \frac{\sigma_{xLoGyLoG} + C_3}{\sigma_{xLoG}\sigma_{yLoG} + C_3} \tag{47}$$

where C_3 avoids inestabilities when $\sigma_{gx}\sigma_{gy}$ is close to zero.

C_1, C_2 and C_3 are calculated as follows:

$$C_i = (K_i L)^2 \tag{48}$$

where L is the dynamic range of the pixel values (255 for 8-bit grayscale images), and $K_i \ll 1$ is a constant. In order to simplify the expression $C_3 = C_2/2$ was chosen. As done in SSIM the following parameter settings were chosen: K_1 =0.01; K_2 =0.03.

Finally, the three indexs of 42, 43 and 38 are combined which results in a specific form of the SSIM index:

$$SSIM(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \cdot \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \cdot \frac{\sigma_{xLoGyLoG} + C_3}{\sigma_{xLoG}\sigma_{yLoG} + C_3} \tag{49}$$

After making the simplification $C_3 = C_2/2$:

$$SSIM(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \cdot \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \cdot \frac{2\sigma_{xLoGyLoG} + C_2}{2\sigma_{xLoG}\sigma_{yLoG} + C_2} \tag{50}$$

Finally the mean of SSIM (MSSIM) is taken to have a single overall value of the frame.

$$MSSIM(X,Y) = \frac{1}{L}\sum_{k=1}^{L} SSIM(x_k, y_k) \tag{51}$$

where X and Y are the reference and the distorted frames, respectively; x_k and y_k are the frame contents at the kth local window; and L is the number of local windows of the image.

3 Performance analysis of video quality metrics

In this section, the two video databases used to validate the metrics described in section 2 are presented in detail. Since the TLabs Database is copyrighted, its sequences are not freely available. Therefore it was decided to use a public database, LIVE Database, so that further studies can compare their results with the ones obtained in this thesis. It is also important to compare the results between different databases, so that the prediction consistency across contents of the video quality metric can be stated. In this section, the performance of the video quality metrics will be presented and analyzed. Since some of the video quality metrics are developed by private companies and therefore are not available for everybody and some others need more information, the tests were performed in some cases only with TLabs Database. At the end of the section the limitations of the examined video quality assessment algorithms will be described.

3.1 Description of databases

3.1.1 TLabs Database

The first database which is used for the tests is the TLabs Database, which consists of five different video contents and two different phases: Phase 1+ and Phase 1++. The reference videos for both phases are the same but with different distortion parameters. The clips of this database are copyrighted and therefore not available on the internet. The videos are provided in avi format and in two definitions: SD (720 $\times$ 576) and HD (1920 $\times$ 1080).

The number of frames as well as the number of frames per second is content dependend and therefore also the duration of each clip. A resume of the values is shown in Table 2.

Content	Description	Number of frames	fps	Duration [s]	Lossy case conditions
A	Movie trailer	382	12.07	31.65	uniform 1% - 4%
B	Interview	398	11.72	33.95	uniform 1% - 4%
C	Soccer sequence	400	8.58	46.63	uniform 1% - 4%
D	Movie sequence	384	13.31	28.86	uniform 1% - 4%
E	Music mail	400	14.02	28.53	uniform 1% - 4%

Table 2: Descriptions of TLabs database characteristics.

1. **Subjective data format**

   The subjective ratings were obtained using the same procedure for every subject in Phase 1+ and 1++. The data was collected after playing a previously randomized playlist following a single-stimulus method, ACR. The MOS results obtained were then converted from R-scale to MOS-11 scale. The playlist had to follow some restrictions:

   - No consecutive clips can be of the same content
   - Anchor files (6 anchors * 5 contents = 30 anchor files per resolution) should be in each test session (per resolution)

   In the screen used for the tests no re-scaling was used as well as no blowing up of the videos. The subjective ratings were taken from the 24 test subjects using a rating scale as can be seen in Figure 13.

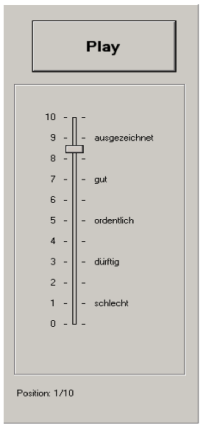


Figure 13: Used scale for the rating of each clip in TLabs Database.

2. **Video characteristics**

   All the snapshots of each video sequence can be seen in the Annex section. In the following, a brief description of each video sequence is presented.

   - **Movie trailer**:
     In this clip lots of different contents like explosion, a fire extinguisher, a running bicycle, etc. appear . Therefore the amount of scene cuts is high (around 7). The sequence has a high level of texture detail and a high amount of motion.

   - **Interview**:
     (interview) This clip is an interview to a girl. There are few scene cuts and medium motion. Since the background does not change much the level of texture detail is low.

   - **Soccer match**:
     This scene is part of a football game. The motion of the sequence is medium and there is a panning camera movement. The level of texture detail is low and there are no scene cuts.

   - **Movie sequence**:
     The movie sequence is a part of wedding in which a few scene cuts appear. The motion as well as the level of texture is high.

   - **U2 music clip**:
     Since it is a music clip there are many scene cuts (between 5 or 6) and a medium amount of motion. The images are dark and different situations like Bono singing or a hand writing in a book are shown. The level of detail is medium.

3.1.2 LIVE Video Quality Database

The second database used for the tests is the so-called LIVE Video Quality Database [29, 30]. The videos of this database are freely available in the Internet [31]. The spatial

resolution of all videos is 768 × 432 and the sequences are providen in YUV format.

In the LIVE Video Quality Database there are ten different reference video contents, which are presented in Table 3.

bs	Blue sky
mc	Mobile and Calendar
pa	Pedestrian Area
pr	Park run
rb	Riverbed
rh	Rushhour
sf	Sunflower
sh	Shields
st	Station
tr	Tractor

Table 3: List of video contents for the LIVE Database.

There are four distortion categories, wireless (four test videos per reference), IP distortions (three test videos per reference), H.264 compression [32] (four test videos per reference), MPEG-2 compression (four test videos per reference), in the LIVE Video Quality Database (plus the reference) and the numbers used for them are showed in Table 4. Distortion strengths were adjusted manually taking care to ensure that the different distorted videos were separated by perceptual levels of distortion.

1	the original reference video
2, 3, 4, 5	Wireless distortions
6, 7, 8	IP distortions
9, 10, 11, 12	H.264 compression
13, 14, 15, 16	MPEG-2 compression

Table 4: List of video distortions for the LIVE Database.

All videos in LIVE Video Quality Database have 250 or 500 frames at a frame rate of 15 fps, which means that some of them have a duration of 16.67 seconds and the others 33.33 seconds. The "bs" sequences have a duration of 14.47 seconds. Sequences "pa", "rb", "rh", "sf", "sh", "st" and "tr" have 250 frames (frame rate of 15fps or 16.67 seconds of video). "mc", "pr" and "sh" have 500 frames (frame rate of 15fps or 33.33 seconds of video).

1. **Subjective data format**

   The subjective ratings were conducted using a a single stimulus procedure and the subjects indicated the quality of the video on a continuous scale. All the videos were viewed by the subjects, including the reference videos to facilitate the computation of different scores using hidden reference removal. In the tests 38 subjects took part and the unreliable subjects were discarded using the procedure specified in ITU-R BT 500.11. In this case 9 of 38 subjects were unreliable and therefore supressed from the provided data. Thus the data is composed by the ratings of the rest 29 valid subjects. The ratings are taken using the Differential Mean Opinion Score (DMOS). Subjects rate the clean version of the video and then the impaired version is substracted, as can be seen in equation 52.

$$DMOS = MOS_{clean} - MOS_{noisy} \tag{52}$$

2. **Video characteristics**

   The snapshots of each video sequence can be seen in the Annex section. In the following, a brief description of each sequence is provided.

   - **Blue sky**:
     The sequence is mainly composed by the leaves and branches of a tree and the sky. In this video the camera motion is slow and rotational it can be observed a high contrast between the dark leaves and the bright sky. The leaves area is highly texturized while the sky area is smooth. The detail level is high.
   - **Mobile and calendar**:
     In this clip appear lots of different objects. It starts with the image of a ship in a calendar and continues with the calendar itself going down over a wall till a small train with little puppets. The camera motion is slow and the amount of movement is low. This close up scene combines a moving calendar with text and a detailed photo of the Vasa ship. After it appears a moving train with colorful toys. The background consists of two types of wallpaper, one brown with details and one yellow with drawn figures. The clip is very detailed and normally demanding.
   - **Pedestrian area**:
     In this movie appears a pedestrian area in which persons and bikes come in the scene. The camera is in a low position and doesn't move at any time while the pedestrians and vehicles move at different speeds close to the camera. The color is constantly varying as the clothes of the persons moving are different. The depth of field is high and a mix of smooth and texture areas is given.
   - **Park run**:
     In this video a man running along a river is seen. The sequence can be divided in to parts, one where a man is running in a park with an umbrella in his hand and the camera is following him at a slow speed, and two when the person stops running and stays, the camera also remains steady. Some texture areas appear like the dark background full of trees and branches, which contrasts with the bright edged area of the snow beside the river. Also the a blurred area apears in the water. The scene is very detailed and demanding.
   - **Riverbed**:
     This scene is formed by a running water of a river. The camera remains steady as the water flows through the screen in a slow speed. The bottom of the water is mainly composed by small stones which do not move and can be seen blurred or textured depending on how the water moves. There is a poor color variation and high brightness areas, this type of scene are very hard to code.
   - **Rushhour**:
     In the sequence can be seen lots of cars and some persons in the background. The camera is fixed and the vehicles and persons move slow or are stopped. The things in the foreground can be clearly seen and mix edge and smooth areas while the background is blurred. The variation of the color is poor while the depth of focus is high.

- **Sunflower**:
  This clip is a very detailed shot, basically composed by a sunflower and a bee, the camera is fixed. While the bee moves slowly over the flower, the antennas and wings meke suddenly fast movements, the global motion is slow. The petals are smooth while the flower stigmas are highly edged and textured. The color of the image remains almost the same along the frames and is mostly a very bright yellow.
- **Shields**:
  This sequence could be also divided in to parts. First of all appears a man with beard and a speckled jacket walking in front of a wall of detailed knight shields and pointing at them, while the camera follows him at a slow motion. Then the man stops and relatevely quick the camera zooms in. In the scene predominates the smooth areas like the wall, but some textured areas can be seen like in the jacket of the man. The shields' structure can be clearly seen since they have noticeable edges.
- **Station**:
  This movie is taken from a bridge to a train station, where many tracks, a train and a woman crossing can be seen. The camera zooms out relatevely fast while the person and the train move slowly. The stones and grass around the tracks define a highly detailed textured area, the tracks themselves describe regular structures and the sky is smooth and dark. It is an evening shot.
- **Tractor**:
  This is the clip with more movement of all. The camera and the tractor move rapidly. Whole sequence contains parts that are very zoomed in and a total view. At the end the camera zoom in blurring some areas of the image. The cultivated land is textured while the parts of the vehicle describe smooth areas. The scene has many colors which points the attention of the viewer.

3.2 Performance of video quality metrics

In this thesis the prediction accuracy will be given by the correlation coefficient of the objective scores to the subjective MOS as a single number value for accuracy performance. This correlation coefficients will be presented in tables together with the Spearman correlation coefficient, the outlier ratios and the root mean square error.

Those were the most relevant parameters used in the VQEGII final report [23] for the methodology for the evaluation of objective model performance. This performance was evaluated with respect to three aspects of their ability to estimate subjective asessment of video quality:

- **Prediction accuracy**, which is the ability to predict the subjective quality ratings with low error. The Pearson linear correlation coefficient between the metrics and the MOS values is the one responsible for this task. It is widely used in the literature as a measure of the strength of linear dependence between two variables.
- **Prediction monotonicity**, which is the degree to which the model's predictions comply with the relative magnitudes of subjective quality ratings. Spearman correlation coefficient measures the extent to which, as one variable increases, the other variable tends to increase, without requiring that increase to be represented by a a linear relationship.

- **Prediction consistency**, which measures the degree to which the model maintains prediction accuracy over the range of video test sequences, i.e., that its response is robust with respect to a variety of video impairments. The root mean square error and the outlier ratio help for this prediction. The outlier ratio is the ratio of "false" scores given by the objective metric to the total number of scores. The "false" scores are the scores that lie outside the interval. The formula for the computation of the outlier ratio of "outlier-points" to total points N is shown in equation 53.

$$oulier\_ratio = \frac{(total\_number\_of\_outliers)}{N} \tag{53}$$

where an outlier is a point for which: $|Q_{error}[i]| > 2{*}metric\_standard\_error[i]$

where Q_{error} is the amount of error for each sample and $metric\_standard\_error[i]$ is the standart error of the metric.

Since the most important parameter is the Pearson correlation, special attention will be paid on it. A score close to 1.0 or -1.0 shows a high prediction accuracy, lower values inform about a low prediction accuracy. A value lower than 0.7 describes a model as weak and lower than 0.5 is considered as unuseable.

Another way to present the results is through scatter-plots, which show the subjective MOS values plotted against the objective scores. In the ideal case all the points would stay around the 45º line. In most of the situations the line will be a fitted curve or line. For points above this line the objective measure predicts a lower quality than derived in the subjective test, points below that line indicate a more optimistic quality prediction.

3.2.1 PSNR

- **TLabs database**

  - Phase 1+:
    Table 5 presents the results for the PSNR video quality metric for the TLabs Database Phase 1+ in the case of considering only the sequences where no packet loss was given. The overall correlation is 0.8051. In Figure 14a it can be seen how the points in the scatter plot are really close to the line, which is an indication of the good performance of PSNR in this kind of transmissions. Considering each of the the contents separately the correlation raises to 0.9858 in the best of the cases. In scenes with few cuts and high level of texture PSNR performs better. Since the worst correlation value is 0.9457 in content "C", it can be concluded that considering the contents separately PSNR has a very satisfactory performance when no packet loss occurs.

    As can be seen in Table 6 PSNR performs weak when using the freezing concealment. The overall correlation value is 0.5833, which is a poor performance. In Figure 14b the scatter plot shows how all the points are spread arround the fitted line. When considering the contents separately the performance does not increase much. Four of the contents have a correlation around 0.7 which stays in the border of the weak performance. The worst correlation value is obtained with content "A", 0.5417, which has many scene cuts and high amount of motion.

    In the slicing concealment scenario the first thing that can be noticed in Table 7 is that the correlation values are better than in the freezing concealment. The

correlation coefficient for all contents is 0.7713, which is a more trustfull value. In Figure 14c it can be seen how all the points stay closer to the line. Thus, PSNR works much bettter when using slicing concealment than when using the freezing one. Special attention should be paid on contents "C" and "D", which have both a correlation coefficient close to 0.9, which is a good performance. Content "A" remains being the one with the worst performance with a value of 0.7268, which is still a good performance. The others, "B" and "E", achieve performances around 0.8. The high amount of motion and the many scene cuts seem to affect the PSNR performance. Also, the concealment causes some mismatch which cannot be measured by PSNR.

The results in Table 8 indicate that the overall performance of PSNR without making any type of classification is weak. The correlation value is of 0.6928 and as it is shown in Figure 14d the points are spread in a cloud around the fitted line. The best correlation is obtained for content "C" and equals 0.8154, which can be evaluated as good. Figure 14d shows the corresponding scatter plot. Values for contents "B", "D" and "E" range between 0.73 and 0.77 which is a good result. The worst performance is obtained in content "A" as in the freezing and slicing concealments. Therefore it can be concluded that when packet losses occur PSNR does not perform well in sequences with lots of scene cuts and high amount of motion and texture detail. Since content "D" and "B" were the best performing before, it can be stated that the number of scene cuts is an important parameter when using PSNR. Since these three contents have few or even no cuts they obtain better results than others with more cuts. Otherwise, for the no-loss case, it performs very well.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.9768	1.0000	0.7500	40.0596
B	0.9785	1.0000	0.5000	33.7792
C	0.9457	0.8000	0.5000	29.2026
D	0.9858	1.0000	0.0000	43.2860
E	0.9783	1.0000	0.5000	31.0221
ALL	0.8051	0.7444	0.3000	35.8738

Table 5: PSNR results for Phase 1+ taking into account only the videos with no packet loss.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.5417	0.4599	0.1905	23.5257
B	0.7092	0.8107	0.0952	19.6015
C	0.6801	0.7400	0.0476	15.8052
D	0.6734	0.7259	0.0714	23.4746
E	0.6983	0.6813	0.0238	17.6026
ALL	0.5833	0.6009	0.0762	20.2405

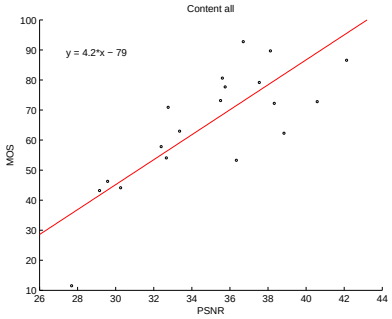
Table 6: PSNR results for Phase 1+ taking into account only the freezing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.7268	0.7626	0.0877	20.5100
B	0.8561	0.8707	0.0702	14.8623
C	0.8721	0.8738	0.0351	14.5465
D	0.9291	0.9456	0.0526	16.2015
E	0.8172	0.8662	0.0351	14.3522
ALL	0.7713	0.7921	0.0596	16.2581

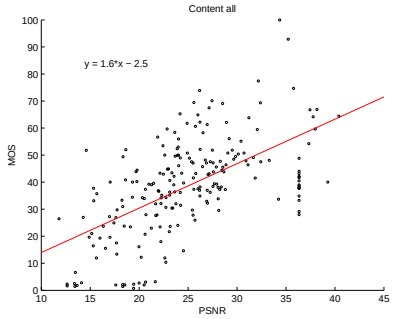
Table 7: PSNR results for Phase 1+ taking into account only the slicing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.5985	0.5857	0.1048	23.2719
B	0.7743	0.7701	0.0762	18.3914
C	0.8154	0.8167	0.0476	16.5705
D	0.7641	0.7599	0.0571	21.7363
E	0.7340	0.7182	0.0476	17.0729
ALL	0.6928	0.6873	0.0686	19.5875

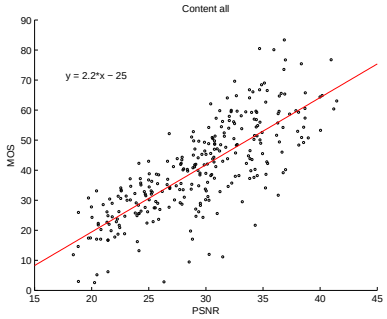
Table 8: PSNR results for Phase 1+ per content for all conditions.



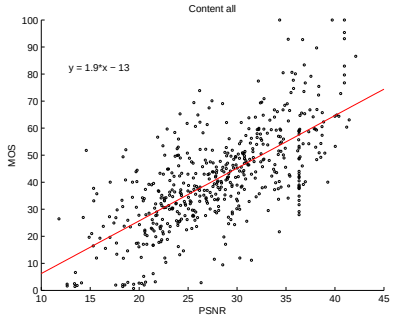
(a) Scatter plot for all the sequences without packet loss.



(b) Scatter plot for all the sequences in the freezing scenario.



(c) Scatter plot for all the sequences in the slicing scenario.



(d) Scatter plot for all the sequences and all the conditions.

Figure 14: Scatter plots for PSNR in TLabs Database Phase 1+.

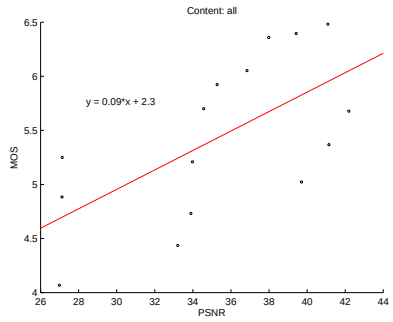
– Phase 1++ for SD resolution:
  Table 9 presents the results for the sequences without packet loss for the TLabs Database Phase 1++. The overall performance, as can be also seen in Figure 15a, is weak, 0.6314. Although in overall the performance is low, when considering the contents separately it can be seen that PSNR works pretty good. The lowest correlation value is for content 'D" and is 0.9277 which is a good result. The rest of the contents have a correlation of 0.97 or even 0.99.

  In Table 10 the correlation coefficients for the freezing scenario are shown. The overall value is 0.3968, which is quite weak. Figure 15b shows how the points are dispersed. For contents "A" and "E" the correlation coefficient is close to 0.1 which is also very weak. Those contents have many scene cuts and have a medium level of motion. Sequence "D" correlates better, 0.5332, while contents "B" and "C" have a correlation value greater than 0.7. With the sequence "C" PSNR achieves the best performance. Both contents, which provide the best performance have few or no scene cuts and the motion is low.

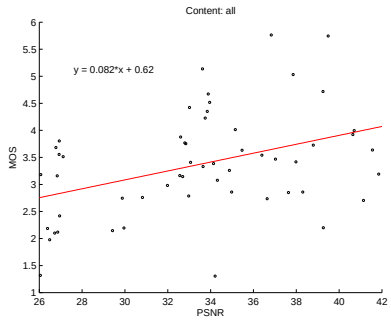
  Table 11 lists the results for PSNR for the slicing scenario of the Phase 1++ data. The correlation values are around 0.9. With content "A" PSNR shows a lower performance, while content "D" seems to fit much better with a correlation value of 0.9441. The overall correlation coefficient is 0.7144, and is quite weak. Figure 15c shows the scatter plot for all the contents in the slicing scenario for the PSNR results in Phase 1++.

In Table 12 the results for all the contents are presented without making any differenciation between the types of concealment or transmission features. The overall correlation is 0.5873, which is categorized as weak. Figure 15d shows the scatter plot with all the points spread in a cloud around the line. PSNR achieves the best performance for content "B" and "D" with correlation coefficients around 0.8. In this case the amount of motion doesn't seem to be an important feature. The lowest performance for Phase 1++ is obtained in content "C", where no scene cuts are given. Content "A" behaves similar than content "C" with a correlation coefficient of 0.6062, but this one has plenty of scene cuts.

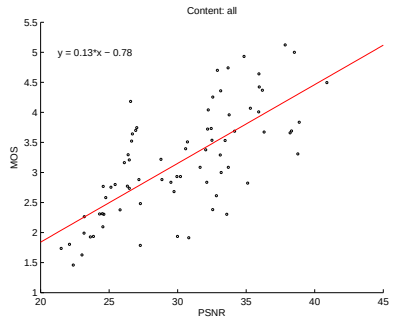
Spearman correlation, outlier ratio and RMSE have similar behaviors than Pearson correlation.



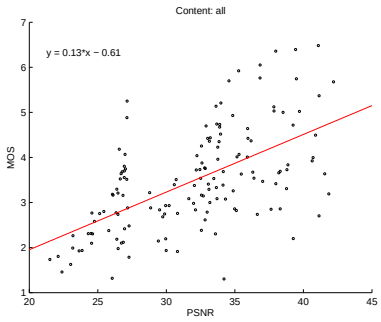
(a) Scatter plot for all the sequences without packet loss.



(b) Scatter plot for all the sequences in the freezing scenario.



(c) Scatter plot for all the sequences in the slicing scenario.



(d) Scatter plot for all the sequences and all the conditions.

Figure 15: Scatter plots for PSNR in TLabs Database Phase 1++ for SD resolution.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.9783	1.0000	1.0000	30.9120
B	0.9794	1.0000	1.0000	28.9972
C	0.9781	1.0000	1.0000	22.3533
D	0.9277	1.0000	1.0000	31.8777
E	0.9983	1.0000	1.0000	35.6667
ALL	0.6314	0.6464	1.0000	30.2799

Table 9: PSNR results for Phase 1++ for SD resolution taking into account only the videos with no packet loss.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.1114	-0.0273	1.0000	31.7236
B	0.7236	0.7727	1.0000	28.8870
C	0.7476	0.6545	1.0000	24.0285
D	0.5332	0.4545	1.0000	30.1925
E	0.1484	0.1182	1.0000	36.0504
ALL	0.3968	0.3768	1.0000	30.4284

Table 10: PSNR results for Phase 1++ for SD resolution taking into account only the freezing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.8878	0.8615	1.0000	28.7811
B	0.9095	0.9321	1.0000	25.9030
C	0.9212	0.9036	1.0000	22.5944
D	0.9441	0.9643	1.0000	26.3519
E	0.8903	0.9036	1.0000	32.3546
ALL	0.7144	0.7255	1.0000	27.3900

Table 11: PSNR results for Phase 1++ for SD resolution taking into account only the slicing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.6062	0.5705	1.0000	30.1503
B	0.7859	0.8749	1.0000	27.3963
C	0.5974	0.6966	1.0000	23.1246
D	0.8175	0.7852	1.0000	28.4619
E	0.6695	0.6532	1.0000	34.1471
ALL	0.5873	0.5968	1.0000	28.8806

Table 12: PSNR results for Phase 1++ for SD resolution per content for all conditions.

– Phase 1++ for HD resolution:
In Table 13 are presented the results for PSNR in TLabs Database Phase 1++ for the HD resolution and the "error-free" scenario. The correlation coefficient for all the contents is 0.2148, which is very low. In Figure 16a is shown the scatter plot. Considering the contents separately the correlation ranges between 0.9186 and 0.9907, which are very good results, but since the amount of samples for each content is low, the validity of these results is limited.

Table 14 shows the results for the freezing concealment scenario. PSNR correlation for all contents is very low again, the coefficient is 0.3630, see Figure 16b. In the figure can be seen two points of the x axis with higher congregation of points and they are when PSNR is around 25 and 29. From 35 on the samples appear more uniformly distributed. Considering the contents separately content "D" correlates much better, 0.8296, than content "A", which correlation is 0.4276.

In the slicing concealment scenario, see Table 15, the correlation coefficient for all the contents is 0.6143, which is a weak performance, see Figure 16c. As can be seen in the figure the number of samples decrease when PSNR increases. The best performance is obtained in content "E" and the correlation is 0.9660. On the other hand the worst result is obtained in content "A", 0.8878, which is also very close to 0.9. Considering the contents separately it can be concluded that PSNR performs good.

As shown in Table 16, where the results for all the contents and conditions are presented, PSNR overall correlation coefficient is 0.4378, which is very low. In Figure 16d can be observed a similar phenomenon than in the freezing concealment scenario. Two points of the x axis have higher congregation of points and they are when PSNR is around 25 and 29. From 35 on the samples appear more uniformly distributed. Considering the contents separately the best result is obtained in content "E", where the correlation is 0.7450, on the other hand the worst is 0.4574 in content "C". Only contents "D" and "E" perform over the 0.7, which is a relative good performance, the others perform weak.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.9779	1.0000	1.0000	30.6224
B	0.9907	1.0000	1.0000	21.8046
C	0.9590	1.0000	1.0000	19.6404
D	0.9186	0.8000	1.0000	30.2070
E	0.9277	0.8000	1.0000	34.2640
ALL	0.2148	0.2180	1.0000	27.8762

Table 13: PSNR results for Phase 1++ for HD resolution taking into account only the videos with no packet loss.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.4276	0.3636	1.0000	33.8339
B	0.8022	0.8727	1.0000	25.3712
C	0.7298	0.5364	1.0000	23.0411
D	0.8296	0.6273	1.0000	32.0965
E	0.6776	0.7455	1.0000	36.0571
ALL	0.3630	0.4506	1.0000	30.4947

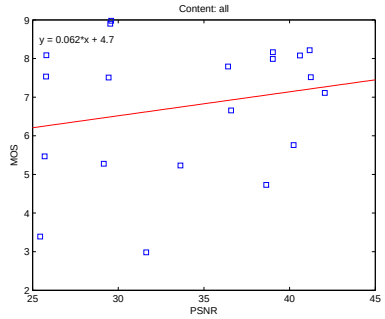
Table 14: PSNR results for Phase 1++ for HD resolution taking into account only the freezing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.8878	0.8964	1.0000	29.5844
B	0.9453	0.9786	1.0000	23.9961
C	0.9428	0.9714	1.0000	21.0841
D	0.8948	0.8500	1.0000	26.1633
E	0.9660	0.9357	1.0000	32.3729
ALL	0.6143	0.5920	1.0000	26.9374

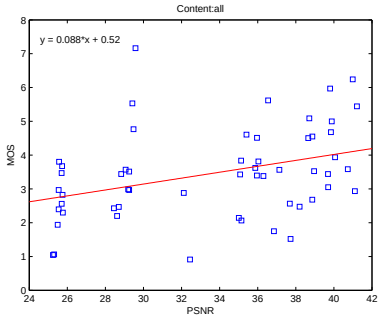
Table 15: PSNR results for Phase 1++ for HD resolution taking into account only the slicing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.6593	0.6142	1.0000	31.3430
B	0.6660	0.8376	1.0000	24.2347
C	0.4574	0.5332	1.0000	21.6417
D	0.7373	0.7781	1.0000	29.0113
E	0.7450	0.8087	1.0000	34.0183
ALL	0.4378	0.5124	1.0000	28.4147

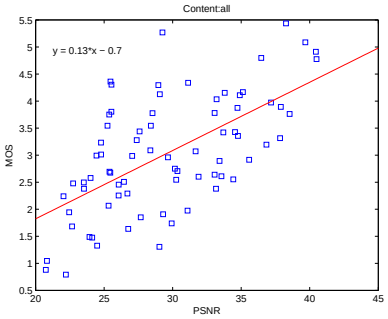
Table 16: PSNR results for Phase 1++ for HD resolution per content for all conditions.



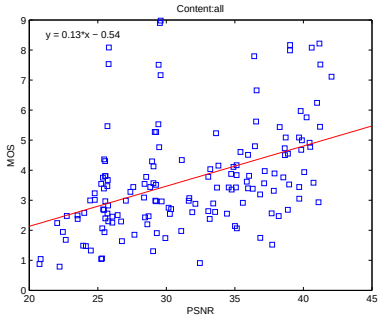
(a) Scatter plot for all the sequences without packet loss.



(b) Scatter plot for all the sequences in the freezing scenario.



(c) Scatter plot for all the sequences in the slicing scenario.



(d) Scatter plot for all the sequences and all the conditions.

Figure 16: Scatter plots for PSNR in TLabs Database Phase 1++ for HD resolution.

- **LIVE database**

  Table 17 contains the PSNR results for the LIVE Database. As shown, the overall performance is low, -0.5507. When considering the contents per separate, the correlation coefficients increases. The video quality of the "tractor", "riverbed" and "rushhour" sequences seem to be well predicted in comparison with the "sunflower" one, which has a correlation of -0.6584. Figure 17 shows the scatter plot for all contents and for all conditions of LIVE Database. As it can be observed, the fitted line has a negative slope, what means that when PSNR increases, the subjective ratings decrease.

  Although the best performance is obtained in the "tractor" sequence, which contains plenty of movement and some quick movements are performed by the camera, PSNR performs also good in low motion sequences with little camera movement, like the "riverbed" and "rushhour" sequences.

  Furthermore, PSNR performs good for most of the contents individually. Only the "sunflower" sequence correlates under -0.7. The "station" sequence is closely under -0.8 while the "park run" clip correlates a bit lower than the "station" one. A reason co'uld be the big edged area similar to the one in the sunflower scene.

  Since the Spearman correlation results are very close to the Pearson correlation the conclusions that could be made are the same. Regarding the prediction consistency, the outlier ratio is almost the same for all contents, between 0.4 and 0.6, and the RMSE between 20 and 30. Therefore it can be concluded that the results are consistent for all contents.

In Table 18 the results are classified by the distortion type. As can be seen, when the clips are compressed using MPEG2, PSNR is not a trustfull method. The correlation is -0.3986 which is considered as unuseable. Also when transmiting the video over wired-internet conections the correlation between PSNR and the subjective results is lower than 0.5. A close value to 0.5 is reached when compressing the videos following H.264 standards. The best result is obtained when the distortions are due to transmission over wireless networks. PSNR performs weak in these cases. Again the correlation coefficients are negative meaning that when PSNR increases, the subjective ratings decrease. Since LIVE Database ratings are taken using the Differential MOS the correlation coefficients are negative. The higher the rating for the impaired version the lower the DMOS.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
pa	-0.8614	-0.8714	0.6000	30.7885
rb	-0.9446	-0.9419	0.6000	25.3775
rh	-0.9187	-0.8964	0.4000	23.3446
tr	-0.9475	-0.9714	0.6000	31.4254
st	-0.7917	-0.6750	0.4000	18.3256
sf	-0.6584	-0.6929	0.5333	20.0509
bs	-0.8680	-0.8179	0.5333	24.2636
sh	-0.8809	-0.8500	0.5333	29.1461
mc	-0.8881	-0.8393	0.6667	31.9342
pr	-0.7534	-0.7846	0.6667	30.3439
ALL	-0.5507	-0.5356	0.5267	26.9093

Table 17: Results with PSNR for all the sequences and conditions in the LIVE Database.

Distortion Type	Pearson corr	Spearman corr	Outlier rat.	RMSE
Wireless	-0.6284	-0.6167	0.7500	33.1865
IP	-0.4717	-0.4145	0.7333	28.7777
H.264	-0.5204	-0.4623	0.3750	23.4624
MPEG 2	-0.3986	-0.3887	0.4250	21.0349

Table 18: Results with PSNR for all the sequences and conditions classified by type of distortion in the LIVE Database.

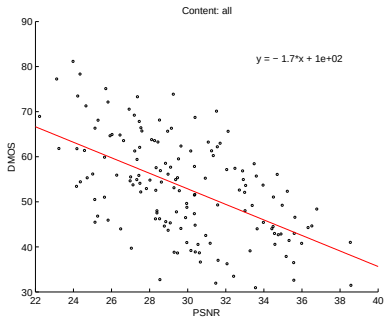


Figure 17: Relation of subjective scores with PSNR for all the sequences and conditions in the LIVE Database.

3.2.2 SSIM

- **TLabs database**

  - Phase 1+:
    In Table 19 the results obtained with the SSIM metric for TLabs Database Phase 1+ are shown. The correlation for all contents is 0.8321, see Figure 18a. SSIM has a good performance when used in no packet loss transmissions. Considering the contents separately it can be seen that they all have a correlation greater than 0.95, which is an excellent performance.

    As shown in Table 20, where the results for all the contents that use the freezing concealment are presented, overall, the metric works good. The correlation is 0.7199. As can be observed in Figure 18b the fitted curve is a second grade polynom, what means that the greater the SSIM value is the higher is the MOS value. Contents "B", "C" and "E" have the highest correlation values, higher than 0.8. On the other hand content "A" has a correlation coefficient of 0.6005. SSIM has a better performance with contents of high level of detail in the freezing concealment scenario.

    The results for the slicing scenario are shown in Table 21. The overall correlation is a bit lower than when using the freezing concealment, 0.6909, Figure 18c. Inspectioning the contents separately can be seen that contents "B", "C" and "D" have a similar behavior, correlating between 0.8063 and 0.8618. Meanwhile, content "A" has a correlation value that equals 0.4856, which means that for this case SSIM correlates very weak. With these results it could be concluded that SSIM depends on the number of scene cuts. The reason for that is that contents "A" and "E", both with low correlation values, are the ones with more scene cuts.

    Results in Table 22 show that the overall performance of SSIM for TLabs Database Phase 1+ for all contents and conditions is weak, the correlation is 0.6468, Figure 18d, which is lower than in all the other categorizations.. Content "C", correlation coefficient is 0.8019 and content "B", 0.7071. Those are the ones that have a correlation value over 0.7. Contents "D" and "E" are close to that value, but still below it. Content "A" correlation is of 0.5102, which is near to the unuseable margin. Therefore SSIM can be rated as a weak metric and it can be concluded that SSIM is scene cut dependent, as explained above.

    As can be seen in Figure 27 all the SSIM plots are fitted with a second order concave curve. This means that the bigger the SSIM predicion is the smaller the subjective rating. One more interesting thing is that most of the points in the plot congregate on the ritght part, meaning that most of the times the SSIM metric is obtaining high values.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.9971	1.0000	1.0000	72.7373
B	0.9914	1.0000	1.0000	66.5511
C	0.9797	0.8000	1.0000	61.8548
D	0.9872	1.0000	0.5000	72.7274
E	0.9601	1.0000	1.0000	68.8828
ALL	0.8321	0.8241	0.9500	68.6730

Table 19: SSIM results for Phase 1+ taking into account only the videos with no packet loss.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.6405	0.5153	0.7381	44.9719
B	0.8139	0.7952	0.6190	40.0041
C	0.8233	0.7638	0.4762	35.6057
D	0.7384	0.7930	0.5000	44.0271
E	0.8394	0.7446	0.6667	42.3445
ALL	0.7199	0.7017	0.5905	41.5259

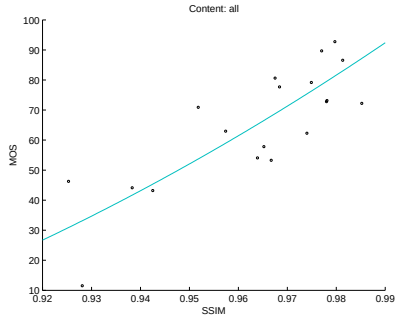
Table 20: SSIM results for Phase 1+ taking into account only the freezing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.4886	0.5797	0.5965	46.7033
B	0.8063	0.8153	0.7895	40.0624
C	0.8279	0.8357	0.8070	40.3315
D	0.8618	0.8876	0.5789	40.6001
E	0.5848	0.7873	0.7368	44.4994
ALL	0.6909	0.7311	0.6947	42.5239

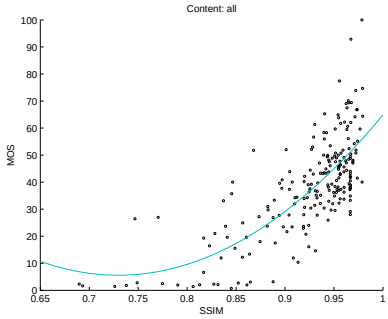
Table 21: SSIM results for Phase 1+ taking into account only the slicing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.5102	0.5422	0.6286	48.0091
B	0.7071	0.7741	0.6667	42.1295
C	0.8019	0.8238	0.5905	40.4757
D	0.6636	0.8073	0.4190	44.6917
E	0.6778	0.7065	0.6857	45.4283
ALL	0.6468	0.7078	0.5943	44.2247

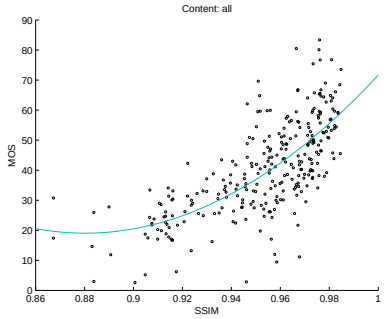
Table 22: SSIM results for Phase 1+ per content for all conditions.



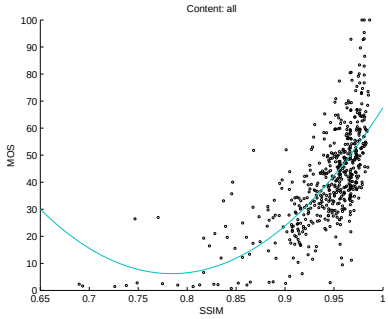
(a) Scatter plot for all the sequences without packet loss.



(b) Scatter plot for all the sequences in the freezing scenario.



(c) Scatter plot for all the sequences in the slicing scenario.



(d) Scatter plot for all the sequences and all the conditions.

Figure 18: Scatter plots for SSIM in TLabs Database Phase 1+.

– Phase 1++ for SD resolution:
  Table 23 presents the results for the "error-free" scenario. As in Phase 1+ the overall correlation is over the 0.7, see Figure 19a, and the contents per separate have correlation values greater than 0.97, which is a quite acceptable result. But the number of samples is still low, therefore the validity of these results is limited.

  In Table 24 are available the results for the SSIM metric when using the freezing concealment. The overall correlation value is low and equals 0.2918, see Figure 19b. And when looking the contents separately it can be seen that the best correlation value is obtained in content "C" and equals 0.3398, which is also very low. SSIM performance in content "D" is even worse, the correlation is -0.0066. This results are that bad that any other conclusion that could be taken would be useless.

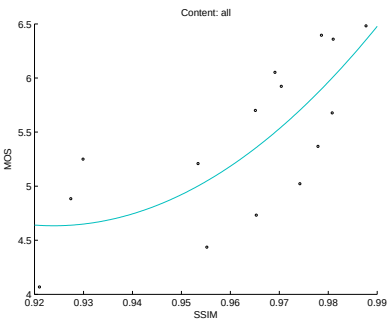
  The results for the slicing concealment are presented in Table 25. The overall correlation is 0.5843, see Figure 19c, which is in the weak margin. What is more important is that when looking at the contents separately the performance of the contents is very good. For instance content "C" and "D" are both above the 0.9 correlation coefficient. On the other hand content "A" is still in the weak range.

  While the correlation value for all contents and conditions is 0.4619, as shown in Table 26 and Figure 19d, which is classified as an unuseable metric, when

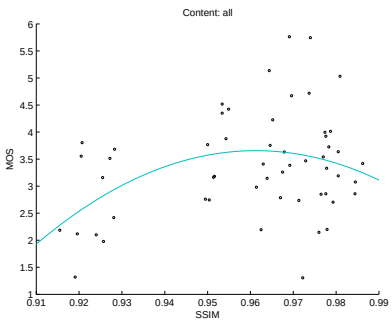
looking at the contents separately, the performance increases. The maximum correlation coefficient is obtained in content "D" and equals 0.6484, which is weak. The lowest value is obtained in content "A", 0.2977, which is very low.

The Spearman correlation performance is similar to the Pearson correlation, therefore the same conclusion can be made. The RMSE is in all cases close to 2 with the exception of the "error-free" scenario where it is around 4, therefore the prediction consistency is almost the same for all the contents.

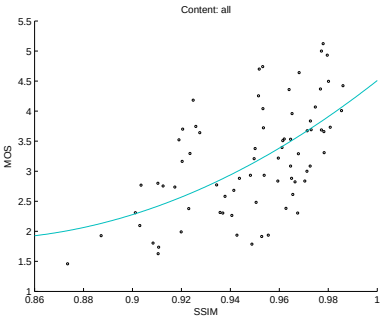
Figure 19 shows all the scatter plots for all the scenarios. As can be seen all figures show a second order polynomic function. Like in Phase 1+ in the "error-free", the slicing and all contents and conditions scenarios the curve is concave, what means that when SSIM increases, MOS values increase quadratically. On the other hand the freezing scenario shows a convex curve, what means that when SSIM values increase, MOS values increase also quadratically but get saturated and start to decrease.



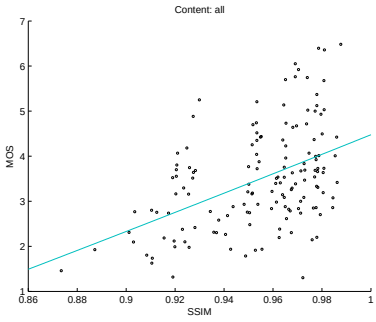
(a) Scatter plot for all the sequences without packet loss.



(b) Scatter plot for all the sequences in the freezing scenario.



(c) Scatter plot for all the sequences in the slicing scenario.



(d) Scatter plot for all the sequences and all the conditions.

Figure 19: Scatter plots for SSIM in TLabs Database Phase 1++ for SD resolution.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.9952	1.0000	1.0000	4.9427
B	0.9778	1.0000	1.0000	4.4371
C	0.9995	1.0000	1.0000	3.8394
D	0.9730	1.0000	1.0000	4.9429
E	0.9993	1.0000	1.0000	4.3864
ALL	0.7184	0.7500	1.0000	4.5284

Table 23: SSIM results for Phase 1++ for SD resolution taking into account only the videos with no packet loss.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	-0.2658	-0.2961	0.7273	2.8635
B	0.2443	0.4091	0.6364	2.6786
C	0.3398	0.3545	0.4545	1.9641
D	-0.0066	-0.0182	0.9091	2.6024
E	-0.2755	-0.2278	0.6364	2.7606
ALL	0.2918	0.2052	0.6727	2.5933

Table 24: SSIM results for Phase 1++ for SD resolution taking into account only the freezing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.5555	0.5559	0.6000	2.8596
B	0.8223	0.8036	0.8000	2.1550
C	0.9177	0.9179	0.7333	2.1550
D	0.9074	0.9179	0.6000	2.3907
E	0.8532	0.8643	0.6667	2.3732
ALL	0.5843	0.6030	0.7067	2.4006

Table 25: SSIM results for Phase 1++ for SD resolution taking into account only the slicing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.2977	0.2869	0.6552	3.1412
B	0.6347	0.6552	0.5172	2.6759
C	0.5780	0.5934	0.5172	2.3217
D	0.6484	0.6348	0.5517	2.8377
E	0.5708	0.5219	0.5862	2.7918
ALL	0.4619	0.4564	0.5655	2.7664

Table 26: SSIM results for Phase 1++ for SD resolution per content for all conditions.

– Phase 1++ for HD resolution:
  In Table 27 are shown the results for the free of errors transmited sequences. The overall correlation value is 0.2429, which is very low. Figure 20a shows the plot for this scenario. Considering the contents separately the correlation coefficient is always higher than 0.92, but since the number of samples per content is very low the validity of the results is limited.

  Table 28 shows the results for the freezing scenario. The correlation for all contents is very low, 0.2886, see Figure 20b. When considering the contents separately it can be seen that except for content "B", which has a correlation coefficient of 0.7640, all of them have a very low correlation value, content "A" the lowest, -0.0090. The difference between content "B" and the rest is huge.

  The results for the slicing scenario are shown in Table 29. The correlation coefficient for all contents is again very low, 0.3610, see Figure 20c. SSIM performs very good in content "C", where the correlation coefficient is 0.9054. On the other hand content "A" has a weak performance, 0.6992.

  In Table 30 can be seen that the overall performance of SSIM is very low, 0.2650. Figure 20d shows the scatter plot. Considering the contents separately can be observed, that the best performance is obtained in content "E", where the correlation is 0.5909. Contents "B", "D" and "E" have a correlation coefficient over 0.5 but, still, far from a good performance.

  Figure 20 shows the scatter plots for all four scenarios. For the freezing concealment, the slicing concealment and all the contents and conditions scenarios the samples draw three vertical curves. Three points of the x axis have higher congregation of points and they are when SSIM is around 0.91, 0.94 and 0.98. Like happened in PSNR, the SSIM estimations for TLabs Database Phase 1++ for HD resolution are not uniformly distributed, they have "predilection" for some values.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.9907	1.0000	1.0000	6.1590
B	0.9886	1.0000	1.0000	6.8837
C	0.9904	1.0000	0.7500	5.5248
D	0.9747	0.8000	0.7500	6.1732
E	0.9211	0.8000	1.0000	5.4058
ALL	0.2429	0.1714	0.9000	6.0526

Table 27: SSIM results for Phase 1++ for HD resolution taking into account only the videos with no packet loss.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	-0.0090	0.1909	0.5455	3.2228
B	0.7640	0.6575	0.2727	3.1364
C	0.0717	0.0818	0.4545	1.8655
D	0.1853	0.2050	0.3636	2.8974
E	0.0215	0.1185	0.5455	2.7377
ALL	0.2886	0.2934	0.3455	2.8140

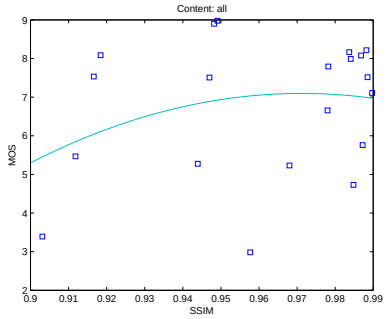
Table 28: SSIM results for Phase 1++ for HD resolution taking into account only the freezing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.6992	0.6821	0.4667	2.4926
B	0.8899	0.9705	0.4000	2.3288
C	0.9054	0.9571	0.5333	2.2223
D	0.8361	0.7679	0.4000	2.1084
E	0.8600	0.9133	0.4000	2.4802
ALL	0.3610	0.4053	0.4533	2.3312

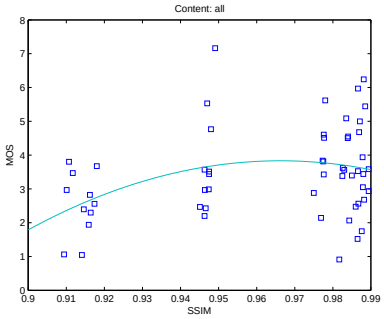
Table 29: SSIM results for Phase 1++ for HD resolution taking into account only the slicing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.4200	0.3891	0.3667	3.4601
B	0.5881	0.7423	0.2333	3.5548
C	0.4329	0.4139	0.1667	2.7956
D	0.5347	0.5784	0.1667	3.2221
E	0.5909	0.6093	0.3333	3.1177
ALL	0.2650	0.3396	0.2467	3.2412

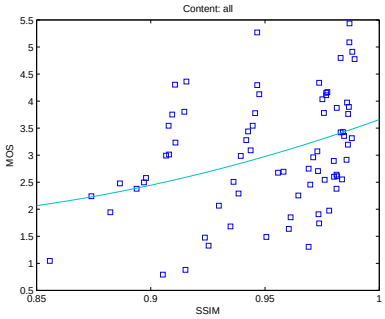
Table 30: SSIM results for Phase 1++ for HD resolution per content for all conditions.



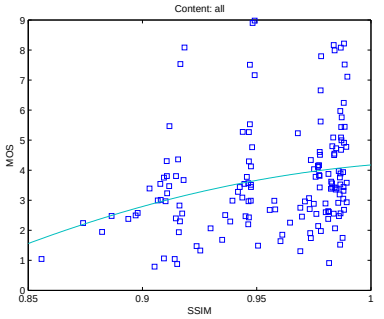
(a) Scatter plot for all the sequences without packet loss.



(b) Scatter plot for all the sequences in the freezing scenario.



(c) Scatter plot for all the sequences in the slicing scenario.



(d) Scatter plot for all the sequences and all the conditions.

Figure 20: Scatter plots for SSIM in TLabs Database Phase 1++ for HD resolution.

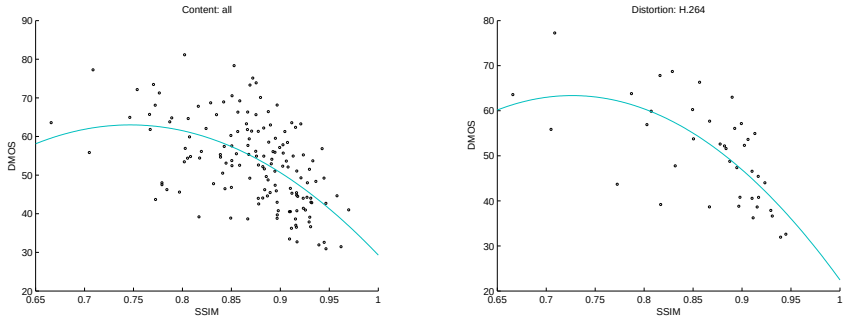
– **LIVE database**

Table 31 presents the SSIM correlation coefficients for LIVE Database. The overall result is poor as the correlation value equals -0.5137. Figure 21a shows the scatter plot for all the contents and condition, like in the freezing concealment scenario of Phase 1++, the curve is convex. It can be also observed, that when each content is analyzed separately SSIM exhibits a very good performance. SSIM has a good performance, -0.9272 is the correlation coefficient, in sequences, where the level of detail is medium or high, like the "tractor", and the "shields" sequences. The "rushhour" and the "park run" sequences correlate around -0.8, which is also a quite acceptable SSIM performance. In the "station", "sunflower", "bluesky" and "mobile and calendar" sequences, where the level of detail is medium, SSIM achieves correlations around -0.7 which is a weak performance. The worst result is obtained in the "riverbed" sequence, where the correlation is -0.5762 and the level of detail is very low. The sequences, for which the camera is steady correlate normally worse than the others. The exception in this case is the "rushhour" sequence, but the "pedestrian area" and the "riverbed" clips, which have no camera movement, correlate worse than the others.

Spearman corerlation results are very close to the Pearson correlation, therefore the conclusions taken are the same as for the Pearson. Regarding the prediction consistency, the RMSE lies between 50 and 60. Therefore it can be said that the consistency of the results is almost the same for all the contents.

In Table 32 the results are listed by the distortion type. The best performance is obtained with the contents coded under the H.264 standard. As is shown in Figure 21b the points do not lie close to the curve, therefore the correlation value is not very high, -0.6147, which is in the weak margin. All the distorsion correlations range between -0.5 and -0.6, what means that the results do not vary much depending on the distortion.

Again, LIVE Database ratings are taken using the Differential MOS, therefore the correlation coefficients are negative. The higher the rating for the impaired version the lower the DMOS.



(a) Relation of subjective scores with SSIM for all the sequences and conditions in the LIVE Database.

(b) Relation of subjective scores with SSIM for all the sequences coded with H.264 in the LIVE Database.

Figure 21: Scatter plots for SSIM in LIVE Database.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
pa	-0.6375	-0.6500	1.0000	59.5688
rb	-0.5762	-0.6023	1.0000	51.4273
rh	-0.8366	-0.8179	1.0000	52.1188
tr	-0.9272	-0.9179	1.0000	56.9301
st	-0.7197	-0.6643	1.0000	49.4844
sf	-0.6925	-0.7286	1.0000	49.4379
bs	-0.7245	-0.7679	1.0000	48.0467
sh	-0.9182	-0.8893	1.0000	54.8615
mc	-0.7002	-0.6500	1.0000	57.8345
pr	-0.7858	-0.7500	1.0000	53.9467
ALL	-0.5137	-0.5515	1.0000	53.4946

Table 31: Results with SSIM for all the sequences and conditions in the LIVE Database.

Distortion Type	Pearson corr	Spearman corr	Outlier rat.	RMSE
Wireless	-0.4813	-0.5448	1.0000	59.1896
IP	-0.5574	-0.4870	1.0000	55.4447
H.264	-0.6147	-0.6801	1.0000	50.8647
MPEG 2	-0.5916	-0.5851	1.0000	48.3226

Table 32: Results with SSIM for all the sequences and conditions classified by type of distortion in the LIVE Database.

3.2.3 VQM

– **TLabs database**

The VQM metric had the restriction that only 15 seconds of each sequence could be computed. Therefore the following decision was taken: first of all the first 15 seconds of the sequence were computed, then the last 15 seconds and finally the metric was ran using the default parameters. As can be seen in the Annex on the VQM tables the results were close to each other. For that reason only the results obtained with the default parameters will be commented and taken as reference.

* Phase 1++ default:
  As shown in Table 33 the performance of the VQM metric in TLabs Database for Phase 1++ in the "error-free" scenario is weak. The correlation is 0.6911. Figure 22a shows the scatter plot. When considering the contents separately it can be seen that the correlation values raise up achieving coefficients above 0.97. Again, since the number of samples for each content is low, the validity of these results is limited.

  Table 34 shows the results for all the sequences in the freezing scenario. As can be seen the overall performance is -0.0257, which is useless. There is no similarity between the metric prediction and the subjective results. Figure 22b shows the scatter plot and as can be observed the slope is almost 0, what means that for same subjective ratings VQM gives different results. For contents "C" and "D" the results are close to the overall. The best performance is obtained with content "E", -0.3888 which is very weak. The unique conclusion that can be made is that when using the freezing concealment, VQM is useless.

  As can be seen in Table 35, VQM seems to perform much better in the slicing concealment scenario. The correlation is 0.6491, which is a weak result, but much better than the one obtained in the freezing concealment scenario. Figure 22c shows the scatter plot. Considering the contents separately all of them have a correlation value close to 0.8 except content "A", which has a correlation of 0.5648. Content "C" correlation is 0.8551. Since it has no scene cuts and medium texture details, it can be concluded that VQM has better performances with contents with few or no scene cuts and medium level of texture details.

  The results for all the contents and conditions are in Table 36. The overall correlation is 0.4383, which defines VQM as a very weak metric. In Figure 22d can be seen all the samples dispersed around the fitted line. Contents "C", "D" and "E" correlation coefficients are close to 0.5, while content "B" has a correlation of 0.6228, which is the best result obtained. On the other hand content "A", has a coefficient of 0.2930 which is the lowest. Again the contents with more cut scenes, motion and texture detail level are the ones wich obtain the lowest performances.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.9998	1.0000	0.6667	1.5831
B	0.9769	1.0000	0.3333	1.4445
C	0.9998	1.0000	1.0000	2.6582
D	0.9841	1.0000	0.3333	1.4737
E	0.9771	1.0000	1.0000	2.8132
ALL	0.6911	0.7214	0.6667	2.0854

Table 33: VQM results for the default seconds of Phase 1++ for SD resolution taking into account only the videos with no packet loss.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	-0.2629	-0.2545	0.6364	3.9276
B	0.2389	0.4182	0.6364	2.9930
C	0.0512	0.0909	1.0000	4.4290
D	-0.1076	-0.1364	0.9091	3.7473
E	-0.3888	-0.2182	0.9091	4.6727
ALL	-0.0257	-0.0118	0.7818	3.9969

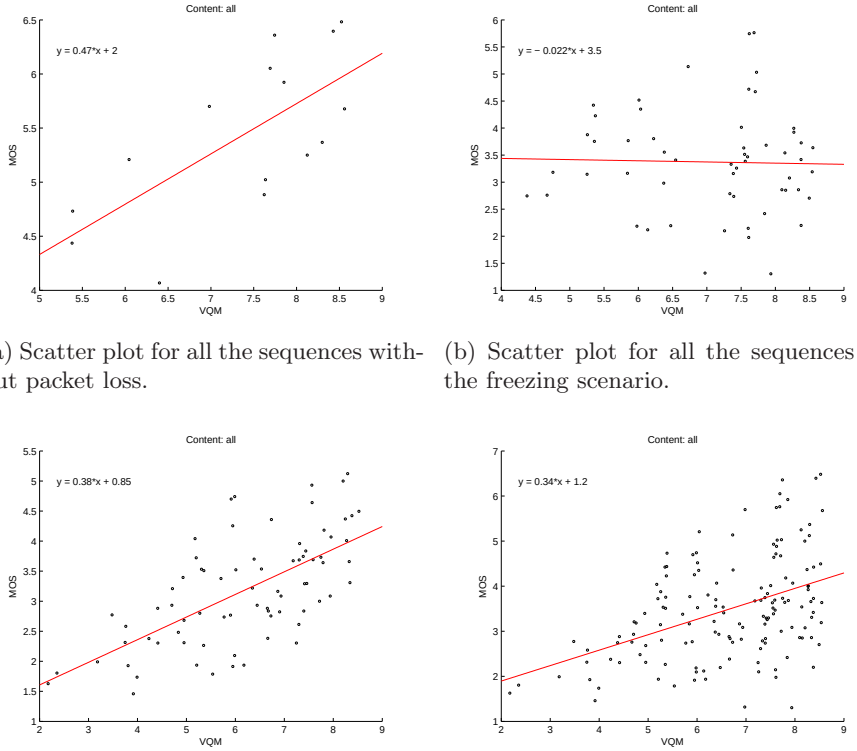
Table 34: VQM results for the default seconds of Phase 1++ for SD resolution and taking into account only the freezing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.5648	0.5719	0.7333	3.0753
B	0.7615	0.7857	0.6000	2.2754
C	0.8551	0.8250	1.0000	3.2729
D	0.7964	0.8143	0.6000	2.8776
E	0.7878	0.7321	1.0000	4.2846
ALL	0.6491	0.6460	0.7867	3.2244

Table 35: VQM results for the default seconds of Phase 1++ for SD resolution and taking into account only the slicing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.2930	0.3227	0.5517	3.3170
B	0.6228	0.6626	0.4138	2.5083
C	0.5259	0.4975	0.9655	3.7030
D	0.5331	0.5507	0.5172	3.1359
E	0.4751	0.4182	0.9655	4.3123
ALL	0.4383	0.4211	0.7034	3.4478

Table 36: VQM results for the default seconds of Phase 1++ for SD resolution per content for all conditions.



(a) Scatter plot for all the sequences without packet loss.

(b) Scatter plot for all the sequences in the freezing scenario.

(c) Scatter plot for all the sequences in the slicing scenario.

(d) Scatter plot for all the sequences and all the conditions.

Figure 22: Scatter plots for VQM in TLabs Database Phase 1++ for SD resolution.

3.2.4 VQuad

– **TLabs database**

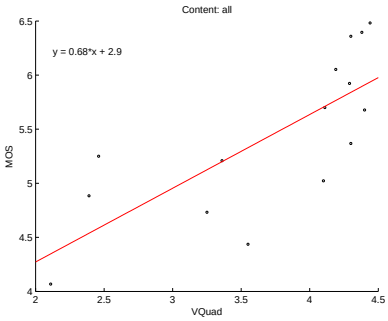
∗ Phase 1++:
In Table 37 are the results for the "error-free" sequences of VQuad video quality metric in TLabs Database Phase 1++. The correlation coefficient for all contents is of 0.7554 which is a good performance. In Figure 23a can be seen the scatter plot. When considering the contents separately it can be noticed that the worst result equals 0.9875 and it is in content "E". All the contents, separately, have an excellent performance.

The results shown in Table 38 are the ones for the freezing scenario. The overall correlation is weak, the value is 0.6859. The scatter plot is in Figure 23b. When considering each content separately the results for contents "B" and "D" perform over 0.9. Specially content "B" has a correlation coefficient of 0.9322. Content "E" has a correlation of 0.5052, which in comparison with the other contents is very low.

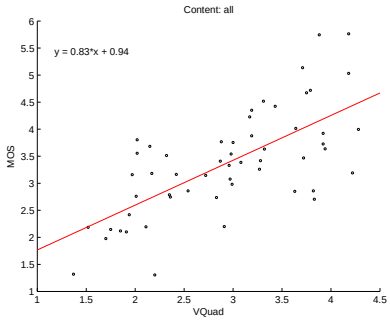
As shown in Table 39 in the slicing scenario the performance of VQuad increases. The correlation value for all the contents together is 0.8089, therefore can be concluded that VQuad performs pretty good with that type of concealment. Figure 23c shows the scatter plot and as can be seen the slope is 0.93, which is close to the ideal case. Four of the contents have a correlation coefficient greater than 0.9 achieving its maximum with

content “A” and a value of 0.9312. On the other hand content “C” is the worst performing one, 0.8363. All the contents correlation coefficients are very close to each other.

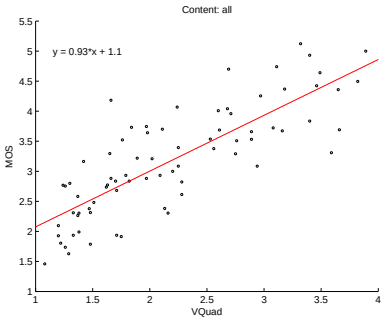
Table 40 presents the results for all the contents and conditions. The overall correlation coefficient is 0.7677, which is a good performance, see Figure 23d, where the slope of the linear fitting is again 0.93, close to the ideal case. The best result is obtained in content “B”, 0.9266. On the other hand the worst performance is obained in content “C” and the correlation coefficient is 0.7402. Contents “B” and “D” have results close to the 0.9 while the others lie close to 0.75. Since contents “B” and “D” have few scene cuts and “C”, the worst performing content, has no scene cuts, it can not be concluded that the amount of scene cuts is a determinant parameter on the performance of VQuad.



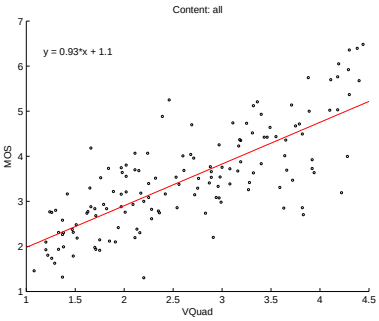
(a) Scatter plot for all the sequences without packet loss.



(b) Scatter plot for all the sequences in the freezing scenario.



(c) Scatter plot for all the sequences in the slicing scenario.



(d) Scatter plot for all the sequences and all the conditions.

Figure 23: Scatter plots for VQuad in TLabs Database Phase 1++ for SD resolution.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.9941	1.0000	1.0000	1.9104
B	0.9954	1.0000	0.3333	1.4118
C	0.9932	1.0000	1.0000	2.4385
D	0.9990	1.0000	0.6667	1.8801
E	0.9875	1.0000	1.0000	1.0991
ALL	0.7554	0.8275	0.7333	1.8073

Table 37: VQuad results for Phase 1++ for SD resolution taking into account only the videos with no packet loss.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.6603	0.4636	0.0000	0.8959
B	0.9322	0.9182	0.0000	0.7158
C	0.8367	0.8656	0.0909	1.0262
D	0.9005	0.8636	0.0000	0.6168
E	0.5052	0.5455	0.0000	0.8541
ALL	0.6859	0.6718	0.0000	0.8340

Table 38: VQuad results for Phase 1++ for SD resolution taking into account only the freezing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.9312	0.9062	0.0000	1.3012
B	0.9228	0.9464	0.0000	0.9347
C	0.8363	0.8811	0.4667	1.5089
D	0.9208	0.8964	0.0000	1.0847
E	0.9217	0.9455	0.0000	0.5191
ALL	0.8089	0.8156	0.0667	1.1216

Table 39: VQuad results for Phase 1++ for SD resolution taking into account only the slicing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.7697	0.7418	0.0000	1.2481
B	0.9266	0.9176	0.0000	0.9233
C	0.7402	0.7376	0.1379	1.4806
D	0.8998	0.8739	0.0000	1.0576
E	0.7677	0.7920	0.0000	0.7356
ALL	0.7691	0.7587	0.0207	1.1191

Table 40: VQuad results for Phase 1++ for SD resolution per content and for all conditions.

3.2.5 MCE

Since the MCE metric needs information of the lost packets, the "error-free" scenario is not presented in this section.

- **TLabs database**
  - Phase 1++ for SD resolution:
    In Table 41 are shown the results for the MCE metric for TLabs Database Phase 1++ in the freezing concealment scenario. The correlation coefficient for all the contents is -0.6875, which is a weak result. In Figure 24a is shown the scatter plot. The best result is obtained in content "B" and the correlation is -0.8217. This content has few scene cuts, medium motion and a low level of texture. Contents "C" and "D" are not far from content "B" performance. On the other hand content "E" correlates close to -0.7 and content "A" correlation is -0.6627. Both contents are the ones with the highest amount of scene cuts.

    As shown in Table 42, where the results for the slicing concealment scenario are presented, the slicing concealment gives more uniform results than the freezing one. The correlation for all contents, see Figure 24b, is -0.8118. Contents "A", "B", "C" and "D" have a correlation between -0.8519 and -0.8822. On the other hand content "E" is highlighting for its lower performance, the correlation is -0.7875.

    Table 43 shows the results for all contents and conditions. The overall correlation coefficient is -0.6221, which is weak, see Figure 24c. Considering the contents separately, the best performance is obtained in content "D" and equals -0,7346. Content "C", which has no scene cuts, is the worst predicted and the correlation is -0.5525. Contents "A", "B" and "D" correlation is around -0.65.

    As can be seen in Figure 24 all the MCE plots are fitted with a second order concave curve. This means that the bigger the MCE predicion is the smaller the subjective rating. One more interesting thing is that most of the points in the plot congregate on the left part, meaning that most of the times the MCE metric is obtaining low values.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	-0.6627	-0.7818	1.0000	1391.9905
B	-0.8217	-0.8182	1.0000	1638.6958
C	-0.7889	-0.8061	1.0000	1722.1804
D	-0.7770	-0.7455	1.0000	1275.9261
E	-0.7238	-0.8545	1.0000	2067.8530
ALL	-0.6875	-0.7797	1.0000	1642.7337

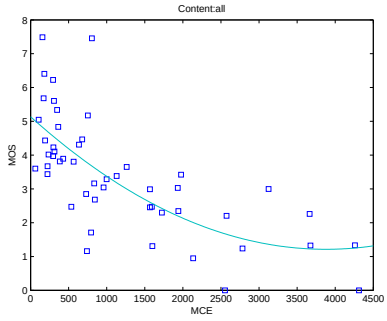
Table 41: MCE results for Phase 1++ for SD resolution taking into account only the freezing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	-0.8822	-0.8823	1.0000	9547.0824
B	-0.8519	-0.9121	1.0000	12966.0083
C	-0.8715	-0.7802	1.0000	13665.6143
D	-0.8549	-0.9165	1.0000	10159.9253
E	-0.7875	-0.9560	1.0000	13243.7477
ALL	-0.8118	-0.8990	1.0000	12038.5556

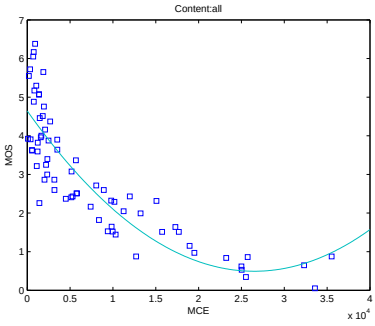
Table 42: MCE results for Phase 1++ for SD resolution taking into account only the slicing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	-0.6312	-0.6876	1.0000	7101.1156
B	-0.6849	-0.8051	1.0000	9597.5579
C	-0.5525	-0.5603	1.0000	10128.0566
D	-0.7346	-0.7518	1.0000	7535.7532
E	-0.6408	-0.8325	1.0000	9850.8035
ALL	-0.6221	-0.7192	1.0000	8932.4397

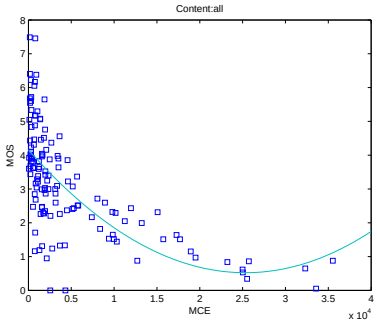
Table 43: MCE results for Phase 1++ for SD resolution per content and for all conditions.



(a) Scatter plot for all the sequences in the freezing scenario.



(b) Scatter plot for all the sequences in the slicing scenario.



(c) Scatter plot for all the sequences and all the conditions.

Figure 24: Scatter plots for MCE in TLabs Database Phase 1++ for SD resolution.

* Phase 1++ for HD resolution:
  In Table 44 are the results of the MCE metric for the freezing concealment scenario. As can be observed the correlation for all contents is -0.6687, which is a weak performance, see Figure 25a. The best performance is obtained in content "A" and the correlation value is -0.8281. Content "D" correlation coefficient is higher than 0.8. The poorest performance is the one obtained in content "B", -0.6327.

  As shown in Table 45, where the results for the slicing concealment scenario are, MCE performs good when using the slicing concealment. The correlation for all contents is -0.8222, see Figure 25b. Content "B" correlation coefficient is -0.8921, closely followed by "C" and "D". On the other hand contents "A" and "E" have correlation coefficients close to 0.8.

  Table 46 presents the results for the MCE for all contents and conditions. As can be seen in Figure 25c the overall performance of the metric is weak, -0.5970. Content "B" is the one which obtains the best performance with a correlation coefficient of -0.6762. Contents "A", "D" and "E" have similar results, around -0.63, while content "C" has a very poor correlation value, -0.4940.

  As can be seen in Figure 25 the MCE metric accumulates most of its predictions on left side of the plot. MCE gives more low than high values. Also as happened in the SD resolution the MCE plots are fitted with a second

order concave curve, what means that the bigger the MCE prediction is, the smaller the subjective rating.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	-0.8281	-0.8545	1.0000	4908.7491
B	-0.6327	-0.7909	1.0000	6442.3347
C	-0.7893	-0.7000	1.0000	6761.1225
D	-0.8032	-0.7364	1.0000	4061.6661
E	-0.7077	-0.8091	1.0000	7146.4761
ALL	-0.6687	-0.7370	1.0000	5981.3297

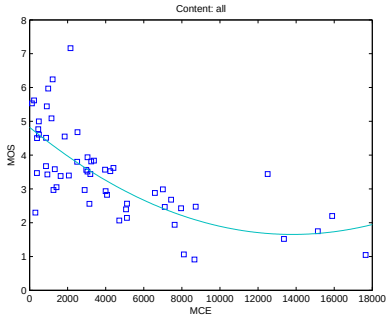
Table 44: MCE results for Phase 1++ for HD resolution taking into account only the freezing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	-0.7985	-0.9536	1.0000	55139.8156
B	-0.8921	-0.9571	1.0000	83545.3795
C	-0.8824	-0.9750	1.0000	87449.0780
D	-0.8610	-0.8607	1.0000	58953.4486
E	-0.8238	-0.9429	1.0000	81047.0670
ALL	-0.8222	-0.9209	1.0000	74446.9034

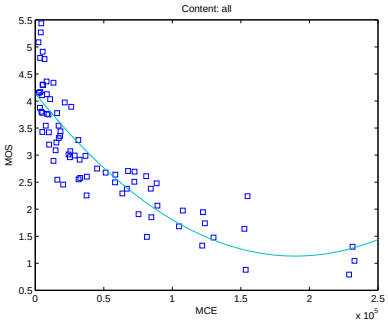
Table 45: MCE results for Phase 1++ for HD resolution taking into account only the slicing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	-0.6208	-0.8174	1.0000	42003.2247
B	-0.6762	-0.7343	1.0000	63595.4755
C	-0.4940	-0.4207	1.0000	66567.7674
D	-0.6398	-0.7461	1.0000	44856.2254
E	-0.6497	-0.7497	1.0000	61734.9170
ALL	-0.5970	-0.6941	1.0000	56680.1623

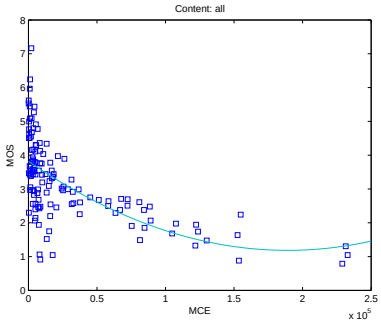
Table 46: MCE results for Phase 1++ for HD resolution per content and for all conditions.



(a) Scatter plot for all the sequences in the freezing scenario.



(b) Scatter plot for all the sequences in the slicing scenario.



(c) Scatter plot for all the sequences and all the conditions.

Figure 25: Scatter plots for MCE in TLabs Database Phase 1++ for HD resolution.

3.2.6 SSIM+

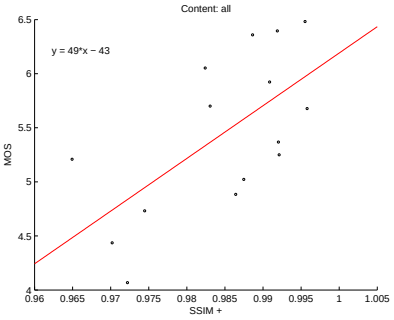
– Phase 1++:
Table 47 presents the results for the "error-free" scenario. The overall correlation is 0.7379, see Figure 26a, and the contents per separate have correlation values greater than 0.97, which is a quite acceptable result. But the number of samples is still low, therefore the validity of these results is limited.

In Table 48 are available the results for the SSIM+ metric when using the freezing concealment. The overall correlation value is low and equals 0.4227, see Figure 26b. And when looking the contents separately it can be seen that the best correlation value is obtained in content "E" and equals 0.4637, which is also low. SSIM+ performance in content "D" is even worse, the correlation is 0.0833. This results are that bad that any other conclusion that could be taken would be useless.

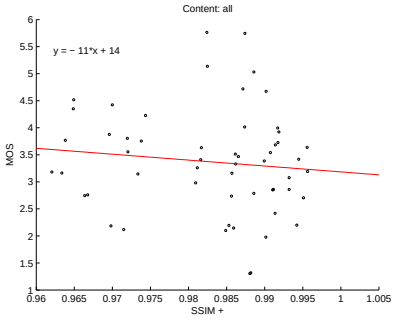
The results for the slicing concealment are presented in Table 49. The overall correlation is 0.6736, see Figure 26c, which is in the weak margin. What is more important is that when looking at the contents separately the performance of the contents improves. For instance content "C" correlation coefficient is 0.9321. On the other hand content "A" is still in the weak range.

While the correlation value for all contents and conditions is 0.5629, as shown in Table 50 and Figure 26d, which is classified as a weak metric, when looking

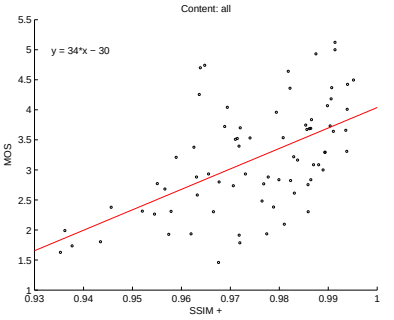
at the contents separately, in some of them the performance increases. The maximum correlation coefficient is obtained in content "D" and equals 0.6171, which is weak. The lowest value is obtained in content "A", 0.3429, which is very low.



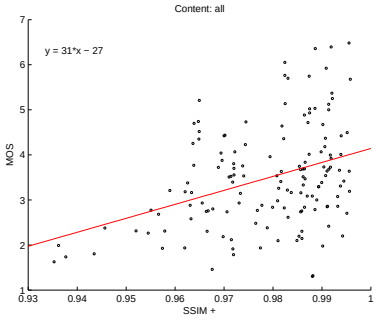
(a) Scatter plot for all the sequences without packet loss.



(b) Scatter plot for all the sequences in the freezing scenario.



(c) Scatter plot for all the sequences in the slicing scenario.



(d) Scatter plot for all the sequences and all the conditions.

Figure 26: Scatter plots for SSIM+ in TLabs Database Phase 1++ for SD resolution.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.9982	1.0000	1.0000	4.9300
B	0.9773	1.0000	1.0000	4.4192
C	0.9998	1.0000	1.0000	3.7818
D	0.9681	1.0000	1.0000	4.9349
E	0.9995	1.0000	1.0000	4.3723
ALL	0.7379	0.5929	1.0000	4.5079

Table 47: SSIM+ results for Phase 1++ for SD resolution taking into account only the videos with no packet loss.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.3443	-0.2455	0.7273	2.8515
B	0.2218	0.2727	0.6364	2.6613
C	0.3042	-0.0455	0.4545	1.9117
D	0.0833	-0.0273	0.9091	2.5939
E	0.4637	-0.2364	0.6364	2.7472
ALL	0.4227	-0.1056	0.6727	2.5746

Table 48: SSIM+ results for Phase 1++ for SD resolution taking into account only the freezing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.6118	0.4004	0.6000	2.8460
B	0.8436	0.6714	0.7333	2.1354
C	0.9321	0.6643	0.7333	2.0961
D	0.9209	0.8893	0.6000	2.3809
E	0.8941	0.7464	0.6667	2.3587
ALL	0.6736	0.5701	0.6667	2.3785

Table 49: SSIM+ results for Phase 1++ for SD resolution taking into account only the slicing concealment.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.3429	0.2158	0.6207	3.1285
B	0.6950	0.5384	0.5172	2.6581
C	0.6267	0.3576	0.5172	2.2668
D	0.6811	0.6020	0.5517	2.8290
E	0.6214	0.4222	0.5862	2.7781
ALL	0.5629	0.3362	0.5517	2.7464

Table 50: SSIM+ results for Phase 1++ for SD resolution per content for all conditions.

– **LIVE database**

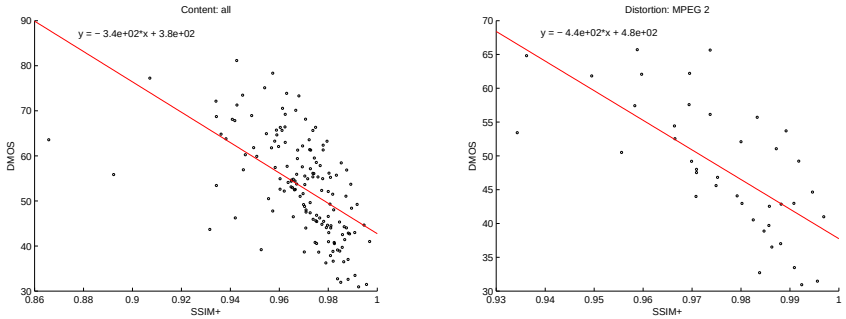
The results for SSIM+ metric tested in LIVE Database are shown in Table 51. The correlation for all contents and conditions is low, and its coefficient is -0.5656. The scatter plot is shown in Figure 27a and as can be seen most of the samples are on the right side of the plots. As happened in SSIM, SSIM+ values are most of the times high. Considering the contents separately can be seen that some of them perform better than others. For instance the "shields" sequence correlation coefficient is -0.9414, which is a very good performance. On the other hand the "riverbed" sequence correlation is -0.4622, which is very weak.

Depending on some of the features of the contents SSIM+ has higher or lower correlation values. Sequences with lower detail level like the "pedestrian area", the "station" and the "mobile and calender", have lower correlation values, around the 0.65. The "sunflower" sequence, is a high detailed one but the correlation coefficient is low. Since the worst correlation values are obtained in sequences, which were filmed with a steady camera, it could be concluded that this parameter influences negatively in the SSIM+ prediction.

Since the Spearman corerlation results are very close to the Pearson correlation the conclusions that could made are the same. Regarding the prediction consistency, the RMSE lies between 45 and 55. Therefore it can be concluded that the consistency of the results is close for all the contents.

In Table 52 the results are classified by the distortion type. All the distortions show almost the same behavior with a correlation coefficient between -0.63 and -0.69. The scatter plot for the best performing distortion, MPEG2, is shown in Figure 27b. Since SSIM+ has the same response to all distortions types and its correlation are under -0.7, it can be concluded that SSIM+ is a weak metric in front of the different distortions.

LIVE Database ratings are taken using the Differential MOS because of that the correlation coefficients are negative. The higher the rating for the impaired version the lower the DMOS.



(a) Relation of subjective scores with SSIM+ for all the sequences and conditions in the LIVE Database.

(b) Relation of subjective scores with SSIM+ for all the sequences coded with MPEG2 in the LIVE Database.

Figure 27: Scatter plots for SSIM+ in LIVE Database.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
pa	-0.5954	-0.6286	1.0000	59.4899
rb	-0.4622	-0.4293	1.0000	51.2621
rh	-0.8088	-0.7793	1.0000	52.0581
tr	-0.8187	-0.7703	1.0000	56.7889
st	-0.6520	-0.6143	1.0000	49.4030
sf	-0.6364	-0.6571	1.0000	49.3763
bs	-0.8014	-0.8893	1.0000	47.9655
sh	-0.9414	-0.9500	1.0000	54.7598
mc	-0.6813	-0.6393	1.0000	57.7414
pr	-0.7898	-0.7786	1.0000	53.8241
ALL	-0.5656	-0.6623	1.0000	53.3957

Table 51: Results with SSIM+ for all the sequences and conditions in the LIVE Database.

Type	Pearson corr	Spearman corr	Outlier rat.	RMSE
Wireless	-0.6348	-0.6568	1.0000	59.0901
IP	-0.6552	-0.5613	1.0000	55.3510
H.264	-0.6247	-0.7313	1.0000	50.7660
MPEG 2	-0.6894	-0.7124	1.0000	48.2181

Table 52: Results with SSIM+ for all the sequences and conditions classified by type of distortion in the LIVE Database.

3.3 Analysis of video quality assessment algorithms

In this section, the performance of the examined video quality metrics will be discussed for both resolutions and for each phase of each database. A detailed analysis of the efficiency of each metric will be presented and their advantages and disadvantages for the different types of scenarios will be analyzed. As in the previous section, the TLabs Database results will be classified by the type of concealment, freezing or slicing, by the transmission errors and all conditions together.

In the following tables, the following notation is used: Pearson stays for correlation, Lev. for level, Cont. for content, H for high, M for medium, L for low and 0 for none. Also Cuts will mean scene cuts, CM camera motion, TD texture detail and SM scene movement.

3.3.1 Analysis of video quality assessment algorithms for TLabs Database Phase 1+

Table 53 shows the comparison between the tested metrics, PSNR and SSIM, with the SD TLabs Database Phase 1+ and the level of scene cuts, motion and texture detail for the "error-free" scenario. SSIM shows a better performance than PSNR. Since the samples for each content are few, any conclusion about the dependencies can not be made.

Table 54 presents the results for the freezing concealment scenario. It can be observed that SSIM has a good performance while PSNR has a weak one. Since the lowest performances are obtained in content "A" and the best are obtained in content "E", it can be concluded that both metrics do not depend on the number of scene cuts. It can be also stated that PSNR is more camera motion dependent than SSIM. It seems to be a relation

between the level of motion and the correlation coefficients: the more motion the lower the correlation coefficient. PSNR is also more texture detail dependent than SSIM, in this case the increase of texture detail means a lower correlation between the metric and the subjective ratings. It is interesting to notice that SSIM correlation coefficients are higher that the PSNR ones for each content.

As can be seen in Table 55 in the slicing concealment scenario PSNR performs better than SSIM. Comparing PSNR and SSIM can be seen that the contents have the same order of performance, the worst results are obtained in content "A" and the best in content "D". PSNR correlation coefficients are higher than SSIM for each content in this scenario. Both video quality metrics depend on the amount of scene cuts given in the sequence, the greater the number of scene cuts the worst is the prediction made. On the other hand they do not depend on the amount of motion or on the level of texture detail.

Comparing both, the freezing and the slicing, concealments can be done the following ranking: $PSNR_S > SSIM_F > SSIM_S > PSNR_F$. This means that in case of knowing the type of concealment used, if it is the freezing one SSIM should be used since it performs better than PSNR. In case of using the slicing concealment PSNR gets better results.

The results for all the sequences and conditions are shown in Table 56. Although both video quality metrics perform weak, PSNR makes better estimations than SSIM, which is not an expected result. The order of the performance for each content is almost the same for both metrics. What is more PSNR shows a dependency on the amount of scene cuts, the more scene cuts the lower the performance. An interesting point is that SSIM does not show more depence than PSNR on the level of texture detail, as would be expected since SSIM is based in the structures recognition.

Considering everything mentioned above can be concluded that in case of not knowing the type of concealment used or the packet-loss rate the best performing metric is PSNR. On the other hand, in case of knowing that the transmission is an "error-free" one, SSIM is the best performing metric. It also shows the best predictions since the correlation coefficient is the highest of Phase 1+.

	PSNR	SSIM	Cuts	CM	TD
Content	Pearson corr.	Pearson corr.	Lev.	Lev.	Lev.
A	0.9768	0.9971	H	H	H
B	0.9785	0.9914	L	L	L
C	0.9457	0.9797	0	M	M
D	0.9858	0.9872	L	H	H
E	0.9783	0.9601	H	M	M
ALL	0.8051	0.8321			

Table 53: Analysis for TLabs Database Phase 1+ for the "error-free" SD sequences.

	PSNR	SSIM	Cuts	CM	TD
Content	Pearson corr.	Pearson corr.	Lev.	Lev.	Lev.
A	0.5417	0.6405	H	H	H
B	0.7092	0.8139	L	L	L
C	0.6801	0.8233	0	M	M
D	0.6734	0.7384	L	H	H
E	0.6983	0.8394	H	M	M
ALL	0.5833	0.7199			

Table 54: Analysis for the SD TLabs Database Phase 1+ for the freezing concealment scenario.

	PSNR	SSIM	Cuts	CM	TD
Content	Pearson corr.	Pearson corr.	Lev.	Lev.	Lev.
A	0.7268	0.4886	H	H	H
B	0.8561	0.8063	L	L	L
C	0.8721	0.8279	0	M	M
D	0.9291	0.8618	L	H	H
E	0.8172	0.5848	H	M	M
ALL	0.7713	0.6909			

Table 55: Analysis for the SD TLabs Database Phase 1+ for the slicing concealment scenario.

	PSNR	SSIM	Cuts	CM	TD
Content	Pearson corr.	Pearson corr.	Lev.	Lev.	Lev.
A	0.5985	0.5102	H	H	H
B	0.7743	0.7071	L	L	L
C	0.8154	0.8019	0	M	M
D	0.7641	0.6636	L	H	H
E	0.7340	0.6778	H	M	M
ALL	0.6928	0.6468			

Table 56: Analysis for TLabs Database Phase 1+ for all the SD contents and conditions.

3.3.2 Analysis of video quality assessment algorithms for TLabs Database Phase 1++

- TLabs Database: Phase 1++ for SD resolution
  Table 57 presents the comparison of the different metrics tested in the SD TLabs Database Phase 1++ and the level of scene cuts, motion and texture detail for the "error-free" scenario. When no packets are lost during the transmission VQuad and SSIM+ are the metrics that perform the best. SSIM, VQM and PSNR got lower correlations. Since the amount of samples per content is low, no dependency can be proved.

  Table 58 shows the results for the freezing concealment scenario. In this case MCE and VQuad have almost the same weak performance. Much lower results are obtained by SSIM+, PSNR and SSIM. VQM performance is the worst one. PSNR and MCE have almost the same content performance order. Both depend on the number of scene cuts, the more cuts the lower correlation coefficients, on the other hand SSIM+ performs better when the amount of scene cuts is higher. Only the MCE video quality metric shows a dependence on the amount of motion of the sequence. With low motion sequences the results obtained are better than when a lot of motion is given in the scene. None of the metrics show a dependence on the level of texture detail. In this scenario PSNR, SSIM, SSIM+ and VQM are unuseable metrics.

  In the slicing scenario, Table 59, case again MCE and VQuad achieve similar results, while PSNR, SSIM+, VQM and SSIM improve their performance, but still far from the best ones. PSNR, SSIM and VQM show a dependency on the nunmber of scene cuts while SSIM+ MCE and VQuad not. SSIM and VQM have the same content performance order and PSNR is very close. None of the video quality metrics are motion dependent and MCE shows a relationship between the amount of texture detail and the subjective ratings. The sequences with a higher level of texture detail show higher correlation values. After observing those results it can be stated that MCE and VQuad predict subjective ratings better than the other metrics when using the freezing or slicing concealments. It can be also concluded that the slicing concealment achieves much better predictions than the freezing one.

  When considering the results for all sequences and conditions, see Table 60, it can be observed that VQuad is, by far, the video quality metric that obtains the best results. After it comes MCE, PSNR, SSIM+, SSIM and at last VQM. Again, MCE and PSNR have the same content performance order. SSIM and VQM show a dependency on the number of scene cuts, the more cuts the lower the correlation coefficient. On the other hand no video quality metric shows a dependence on the amount of motion or on the level of texture detail. Once all the results are observed it can be concluded that MCE is roughly better than VQuad when the type of concealment used is known. When not, VQuad achieves better predictions than the other metrics.

	PSNR	SSIM	VQM	VQuad	MCE	SSIM+	Cuts	CM	TD
Content	Pearson	Pearson	Pearson	Pearson	Pearson	Pearson	Lev.	Lev.	Lev.
A	0.9783	0.9952	0.9998	0.9941	-	0.9982	H	H	H
B	0.9794	0.9778	0.9769	0.9954	-	0.9773	L	L	L
C	0.9781	0.9995	0.9998	0.9932	-	0.9998	0	M	M
D	0.9277	0.9730	0.9841	0.9990	-	0.9681	L	H	H
E	0.9983	0.9993	0.9771	0.9875	-	0.9995	H	M	M
ALL	0.6314	0.7184	0.6911	0.7554	-	0.7379			

Table 57: Analysis for TLabs Database Phase 1++ for the "error-free" SD sequences.

	PSNR	SSIM	VQM	VQuad	MCE	SSIM+	Cuts	CM	TD
Content	Pearson	Pearson	Pearson	Pearson	Pearson	Pearson	Lev.	Lev.	Lev.
A	0.1114	-0.2658	-0.2629	0.6603	-0.6627	0.3443	H	H	H
B	0.7236	0.2443	0.2389	0.9322	-0.8217	0.2218	L	L	L
C	0.7476	0.3398	0.0512	0.8367	-0.7889	0.3042	0	M	M
D	0.5332	-0.0066	-0.1076	0.9005	-0.7770	0.0833	L	H	H
E	0.1484	-0.2755	-0.3888	0.5052	-0.7238	0.4637	H	M	M
ALL	0.3968	0.2918	-0.0257	0.6859	-0.6875	0.4227			

Table 58: Analysis for the SD TLabs Database Phase 1++ for the freezing concealment scenario.

	PSNR	SSIM	VQM	VQuad	MCE	SSIM+	Cuts	CM	TD
Content	Pearson	Pearson	Pearson	Pearson	Pearson	Pearson	Lev.	Lev.	Lev.
A	0.8878	0.5555	0.5648	0.9312	-0.8822	0.6118	H	H	H
B	0.9095	0.8223	0.7615	0.9228	-0.8519	0.8436	L	L	L
C	0.9212	0.9177	0.8551	0.8363	-0.8715	0.9321	0	M	M
D	0.9441	0.9074	0.7964	0.9208	-0.8549	0.9209	L	H	H
E	0.8903	0.8532	0.7878	0.9217	-0.7875	0.8941	H	M	M
ALL	0.7144	0.5843	0.6491	0.8089	-0.8118	0.6736			

Table 59: Analysis for the SD TLabs Database Phase 1++ for the slicing concealment scenario.

	PSNR	SSIM	VQM	VQuad	MCE	SSIM+	Cuts	CM	TD
Content	Pearson	Pearson	Pearson	Pearson	Pearson	Pearson	Lev.	Lev.	Lev.
A	0.6062	0.2977	0.2930	0.7697	-0.6312	0.3429	H	H	H
B	0.7859	0.6347	0.6228	0.9266	-0.6849	0.6950	L	L	L
C	0.5974	0.5780	0.5259	0.7402	-0.5525	0.6267	0	M	M
D	0.8175	0.6484	0.5331	0.8998	-0.7346	0.6811	L	H	H
E	0.6695	0.5708	0.4751	0.7677	-0.6408	0.6214	H	M	M
ALL	0.5873	0.4619	0.4383	0.7691	-0.6221	0.5629			

Table 60: Analysis for TLabs Database Phase 1++ for all the SD contents and conditions.

- TLabs Database: Phase 1++ for HD resolution
  In Table 61 are the results for the HD TLabs Database Phase 1++ for the "error-free" sequences. Both metrics, PSNR and SSIM, have very low performance. It can be concluded that they are unuseable for these type of sequences, although for each content individually, they perform very good.

  In Table 62 are shown the results for the freezing concealment scenario. MCE metric has a much better performance than PSNR, which at the same time achieves better results than SSIM. MCE video quality metric is in the weak performance range while PSNR and SSIM are classified as unuseable. PSNR and SSIM have almost the same content performance order, being content "A" the one with the lowest correlation value. PSNR and SSIM show a dependency on the amount of scene cuts. The higher the amount of scene cuts given in a sequence the worse the performance. On the other hand MCE is motion dependent, when motion level of the sequence is high is when the metric correlates better with the subjective ratings. MCE is also dependent on the amount of texture detail. With high levels of detail MCE's prediction gets better.

  In the slicing concealment scenario, see Table 63 MCE also has better results than PSNR and SSIM. The video quality metric ranking taking into account both concealments is this: $MCE_S \gg MCE_F > PSNR_S \gg PSNR_F > SSIM_S > SSIM_F$. In this case MCE shows a relationship between the scene cuts and its performance, the less scene cuts the better results are obtained. SSIM looks like also has this dependency but in a lower measure. MCE, PSNR and SSIM show a high dependency between the motion level (the more motion the lower the performance). PSNR shows a higher dependency than the other video quality metrics on the level of texture detail, when the sequence has a lot of texture information the prediction is worse. An interesting phenomenon is that PSNR achieves better results than SSIM and MCE in each content, but not in the overall. Therefore if the type of content is known and the slicing concealment is used it would be recommended to use PSNR as prediction metric. In all other cases the advice would be to use the MCE video quality metric.

  In Table 64 are presented the results for all sequences and conditions. MCE, again, performs better than PSNR and SSIM. In this case no of the video quality metrics shows any type of dependency with the amount of scene cuts, camera motion or texture detail.

  To conclude, it could be stated that the best performances are obtained when using the slicing concealment and that the MCE video quality metric is the one that always has a better performance with the HD resolution.

	PSNR	SSIM	MCE	Cuts	CM	TD
Content	Pearson	Pearson	Pearson	Lev.	Lev.	Lev.
A	0.9779	0.9907	-	H	H	H
B	0.9907	0.9886	-	L	L	L
C	0.9590	0.9904	-	0	M	M
D	0.9186	0.9747	-	L	H	H
E	0.9277	0.9211	-	H	M	M
ALL	0.2148	0.2429	-			

Table 61: Analysis for TLabs Database Phase 1++ for the "error-free" HD sequences.

	PSNR	SSIM	MCE	Cuts	CM	TD
Content	Pearson	Pearson	Pearson	Lev.	Lev.	Lev.
A	0.4276	-0.0090	-0.8281	H	H	H
B	0.8022	0.7640	-0.6327	L	L	L
C	0.7298	0.0717	-0.7893	0	M	M
D	0.8296	0.1853	-0.8032	L	H	H
E	0.6776	0.0215	-0.7077	H	M	M
ALL	0.3630	0.2886	-0.6687			

Table 62: Analysis for the HD TLabs Database Phase 1++ for the freezing concealment scenario.

	PSNR	SSIM	MCE	Cuts	CM	TD
Content	Pearson	Pearson	Pearson	Lev.	Lev.	Lev.
A	0.8878	0.6992	-0.7985	H	H	H
B	0.9453	0.8899	-0.8921	L	L	L
C	0.9428	0.9054	-0.8824	0	M	M
D	0.8948	0.8361	-0.8610	L	H	H
E	0.9660	0.8600	-0.8238	H	M	M
ALL	0.6143	0.3610	-0.8222			

Table 63: Analysis for the HD TLabs Database Phase 1++ for the slicing concealment scenario.

	PSNR	SSIM	MCE	Cuts	CM	TD
Content	Pearson	Pearson	Pearson	Lev.	Lev.	Lev.
A	0.6593	0.4200	-0.6208	H	H	H
B	0.6660	0.5881	-0.6762	L	L	L
C	0.4574	0.4329	-0.4940	0	M	M
D	0.7373	0.5347	-0.6398	L	H	H
E	0.7450	0.5909	-0.6497	H	M	M
ALL	0.4378	0.2650	-0.5970			

Table 64: Analysis for TLabs Database Phase 1++ for all the HD contents and conditions.

3.3.3 Analysis of video quality assessment algorithms for LIVE Database

The results of the metrics tested with LIVE Database are presented in Table 65. As can be seen SSIM+ has a better performance than PSNR, which at the same time has a better performance than SSIM. All the metrics have a weak correlation value. Since SSIM+ is based on SSIM the order of the performance of the contents is very similar.

In all three video quality metrics can not be demonstrated that a depency between camera motion and scene movement level exists. On the other hand, and as expected, can be seen that exists a little relationship between the amount of texture detail and SSIM+ and SSIM. The correlation coefficients raise up when the level of detail is higher.

In the Table 66 can be seen that SSIM+ has always better results with all types of distortions. SSIM+ shows in all cases the best performances and therefore is the best predicting video quality metric for this database.

Since LIVE Database ratings are taken using the Differential MOS the correlation coefficients are negative. The higher the rating for the impaired version the lower the DMOS.

	PSNR	SSIM	SSIM+	CM	SM	TD
Content	Pearson	Pearson	Pearson	Lev.	Lev.	Lev.
pa	-0.8614	-0.6375	-0.5954	M	M	M
rb	-0.9446	-0.5762	-0.4622	M	L	L
rh	-0.9187	-0.8366	-0.8088	0	L	M
tr	-0.9475	-0.9272	-0.8187	0	H	H
st	-0.7917	-0.7197	-0.6520	L	L	M
sf	-0.6584	-0.6925	-0.6364	M	L	H
bs	-0.8680	-0.7245	-0.8014	0	L	H
sh	-0.8809	-0.9182	-0.9414	0	L	M
mc	-0.8881	-0.7002	-0.6813	L	L	H
pr	-0.7534	-0.7858	-0.7898	L	M	H
ALL	-0.5507	-0.5137	-0.5656			

Table 65: Analysis for LIVE Database for all the sequences and conditions.

	PSNR	SSIM	SSIM+
Dist. Type	Pearson	Pearson	Pearson
Wireless	-0.6284	-0.4813	-0.6348
IP	-0.4717	-0.5574	-0.6552
H.264	-0.5204	-0.6147	-0.6247
MPEG 2	-0.3986	-0.5916	-0.6894

Table 66: Analysis for LIVE Database for the different types of distortions.

3.3.4 Limitations of video quality assessment algorithms

- **Wrong assumptions**

  Error sensitivity methods like PSNR have traditionally been used in analogue systems as a consistent quality metric. However, digital video technology has exposed some limitations of MSE/PSNR. The reason for these limitations lie in the composition of the human visual system (HVS). Error sensitivity methods are based in the functional properties of early stages of the HVS, which is rather complex and highly nonlinear. These bottom-up approaches have found nearly universal acceptance because of its easy implementation and comprehension, but since these models are based on linear operator, the methods must rely on a number of strong assumptions and generalizations. In [13] some of these assumptions are taken into account:

  - **Quality definition problem:**
    As was shown in [33] the correlation between image fidelity and image quality is only moderate. Therefore the definition of images quality stays as an open question: does error visibility mean loss of quality? Some distortions might be clearly visible, but maybe not so objectionable.
  - **Suprathreshold problem:**
    Since the error sensitivity models are based on threshold estimators, which are extracted from models which are specifically designed to estimate threshold at which a simulus is barely visible, it's not realistic to think that this threshold approaches work to characterize perceptual distortions, which are significantly larger than the threshold values. The open question in this case would be: can these models be generalized for the suprathreshold range?
  - **Natural image complexity problem:**
    Since natural images have a highly complicated composition and most psychophysical experiments are conducted using relatively simple patterns the next question appears: can a few simple patterns build a model that can predict the quality of a complex-structured natural image? At this time the answer for this question is still unknown but [34] should facilitate further studies.
  - **Decorrelation problem:**
    When e.g. using a Minkowski metric for spatially pooling errors, it is been assumed that errors at different locations are statistically independent. However in [35, 36] has been shown that exists a strong dependency between intra- and inter- channel wavelet coefficients of natural images.
  - **Cognitive interaction problem:**
    Cognitive understanding of an image influences on the perceived quality. Also, prior information regarding the image content or the attention and fixation taken to watch the scene may also affect in the evaluation of the video quality. But since these effects are difficult to quantify and not well understood, they are not considered in most image quality metrics.

  In [37] other assumptions that are usually misunderstood, when using error sensitivity models, are taken into account. Notice that all of these assumptions are wrong.

  - **Signal fidelity is independent of temporal or spatial relationships between the samples of the original signal.** In other words, if the original

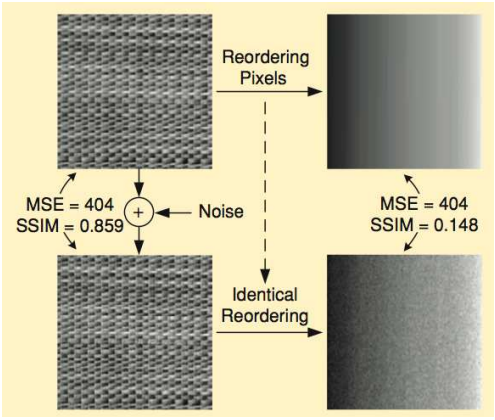


Figure 28: Signal fidelity independence of temporal or spatial relationships.

and distorted signals are randomly reordered in the same way, then the MSE between them will be unchanged.

In Figure 28 can be observed that an original image (top left) is distorted by adding independent white Gaussian noise (bottom left). In the top-right image, the pixels are reordered by sorting pixel intensity values. The same reordering process is applied to the bottom-left image to create the bottom-right image. The MSE between the two left images and between the two right images are the same, but the bottom-right image appears much noisier than the bottom-left image. The perceived visual fidelity of the bottom-right image is much poorer than that of the bottom-left image.

Apparently, this assumption is not realistic when measuring the fidelity of images. Since natural images are highly structured and HVS is highly adapted to structure recognition the ordering of the signal samples carries important and valuable structural information. Therefore MSE/PSNR might not measure well the fidelity in signals which are highly structured.

– **Signal fidelity is independent of any relationship between the original signal and the error signal.** For a given error signal, the MSE remains unchanged, regardless of which original signal it is added to.

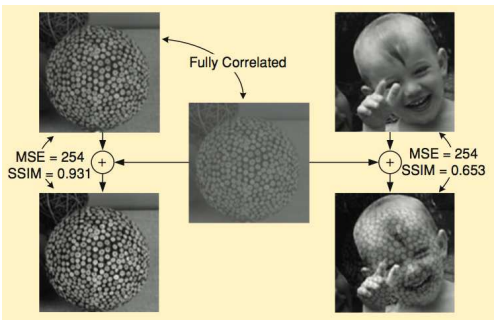


Figure 29: Signal fidelity independence of relationships between the original and distorted signal.

As can be seen in Figure 29 two original images (top left and top right) are distorted by adding the same error image (middle), which is fully correlated with the top-left image. The MSE between the two left images and between the two right images are the same, but the perceived distortion of the bottom-

right image is much stronger than that of the bottom-left image. Clearly, the correlation (and dependency) between the error signal and the underlying image signal significantly affects perceptual image distortion.

– **Signal fidelity is independent of the signs of the error signal samples.**

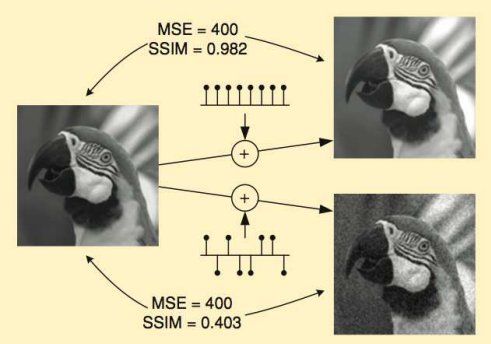


Figure 30: Signal fidelity independence of the signs of the error signal samples.

Figure 30 depicts how an original image (left) is distorted by adding a positive constant (top right) and by adding the same constant, but with random signs (bottom right). The visual fidelity of the two distorted images is drastically different. Yet, the error sensitivity metric (MSE/PSNR) ignores the effect of signs and reports the same fidelity measure for both distorted images.

– **All signal samples are equally important to signal fidelity.**

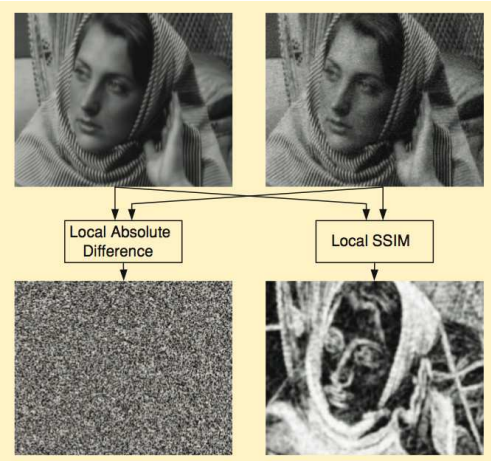


Figure 31: All signal samples are equally important to signal fidelity.

In Figure 31 an original image (top left) is distorted by adding independent white Gaussian noise (top right). The energy distribution of the absolute difference signal (bottom left, enhanced for visibility), is uniform. The perceived noise level is space variant, which is reflected in the SSIM map (bottom right, enhanced for visibility). the distorted image (top right) was created by adding independent white Gaussian noise to the original image (top left). Clearly, the degree of noise-induced visual distortion varies significantly across the spatial coordinates of the image. In particular, the noise in the facial (and other smooth-intensity) regions appears rather severe, yet is visually negligible in other regions containing patterns and textures. The perceived fidelity of the

distorted image varies over space, although the error signal (bottom left) has an uniform energy distribution across space. Since all image pixels are treated equally in the formulation of the error sensitive metrics (MSE/PSNR) , such image content-dependent variations in image fidelity cannot be accounted for.

Another approach to deal with the perceived fidelity problem would be the following. Finding the maximum/minimum SSIM images along an equal-MSE hypersphere in image space.

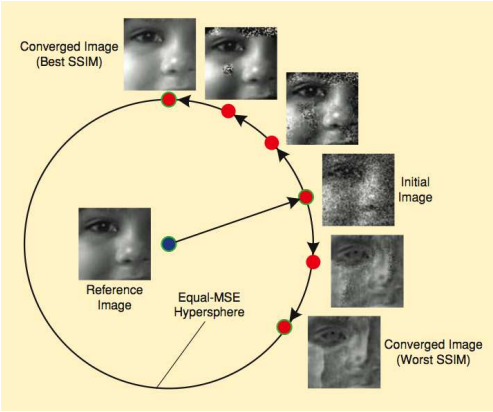


Figure 32: Finding the maximum/minimum SSIM images along the equal-MSE hypersphere in image space.

In Figure 32 the images on the hypersphere have the same MSE/PSNR but are substantially different perceived. The length of the distortion vector does not suffice as a good indication of image fidelity. Apparently, the directions of these vectors are also important.

- **MSE/PSNR are content dependent**

As shown in [38] it is provable that as long as the video content and the codec type are not changed, PSNR is a valid quality measure. However, when the content is changed, correlation between subjective quality and PSNR is highly reduced. Hence PSNR cannot be a reliable method for assessing the video quality across different video contents.

Here is presented some experimental data that demonstrates where and why PSNR can or cannot be used as a quality metric. The sequences used for this experiment are the same as in phase 1++.

The implication is that, within a specified codec and fixed content, the variation of PSNR is a reliable indicator of the variation of quality. Hence, in the context of codec optimization, PSNR can therefore be used as a performance metric as it correlates highly with subjective quality when the content is fixed. For example, PSNR can be used for testing different codec optimization strategies designed to maximize the subjective quality of a specified content.

Content	All	1	2	3	4	5
Correlation	0.6314	0.9783	0.9794	0.9781	0.9277	0.9983

Table 67: Correlation between PSNR and quality per content and across contents.

Table 67 shows that, if PSNR is used across contents, it loses its prediction accuracy. The correlation between PSNR and subjective quality for the data in phase 1++ when

there is no packet loss is only 0,6314. The data in the table 67 shows that, when each content is considered separately, the Pearson correlation between PSNR and subjective quality is well above 0.9, but the value dramatically drops to 0.63 when all data points from the five contents are considered together in the test set. If more sources of content were jointly assessed, the Pearson correlation would be much lower than this.

These results show that PSNR is therefore unreliable as an objective metric for predicting subjective quality. Furthermore, the measured correlation coefficient of 0.63 between the PSNR and subjective quality shown in table 67 is only for five video contents. Adding more contents can only reduce this correlation coefficient further, whereas the individual correlation coefficient for each content remains high. This reiterates the conclusion that PSNR is not a reliable measure of quality across various video contents, but it is reliable within the content itself.

4 Conclusions

4.1 Summary of results

This thesis has conducted research related to objective video quality assessments. In this chapter, the results will be summarized, and then some possible future works, will be suggested.

Perceptual video quality assessment has flourished in recent years. Therefore the research discussed and models proposed in this thesis are important supplements to the existing research work in literature.

The impact of several parameters (scene cuts, camera motion, texture detail, etc.) on objective video quality assessment methods with different databases and resolutions was examined. Then a new metric for video quality prediction, SSIM+, was proposed and tested.

The results obtained in TLabs Database Phase 1+ for the SD resolution show that for all contents and condition PSNR performs better than SSIM. In the"error-free" scenario SSIM shows a better performance than PSNR and in case of knowing the type of concealment used, if it is the freezing one, SSIM performs better than PSNR. In case of using the slicing concealment PSNR gets better results. PSNR shows a dependency on the amount of scene cuts, the more scene cuts the lower the performance. On the other hand SSIM does not show more depence than PSNR on the level of texture detail, as would be expected since SSIM is based in the structures recognition.

Phase 1++ results for SD resolution and for all sequences and conditions, show that VQuad is, by far, the video quality metric that obtains the best performances. After it comes MCE, PSNR, SSIM, VQM and at last SSIM+. SSIM and VQM show a dependency on the number of scene cuts, the more cuts the lower the correlation coefficient. When no packets are lost during the transmission VQuad is the metric that performs the best. Once all results are observed it can be stated that MCE is roughly better than VQuad when the type of concealment used is known. When not, VQuad achieves better predictions than the other metrics. The slicing concealment gets better results than the freezing one for each metric.

In LIVE Database, once again, PSNR performs better than SSIM across all contents. On the other hand SSIM+ outperforms both metrics. In all three video quality metrics could not be demonstrated that a depency between camera motion and scene movement level existed. As expected the correlation coefficients of SSIM and SSIM+ raise up when the level of detail is higher.

A very generic method to increase the prediction accuracy of visual quality metrics was proposed. Results show, that the simple SSIM+ metric, that was built using the described method (and still does nothing more than pure SSIM calculation) outperforms PSNR and SSIM quality metrics. The gap between the result obtained with SSIM+ and pure SSIM was achieved by the introduction of a very simple correction step. While the approach has been tested for videos of LIVE Database in general this method should also work for other databases. SSIM+ also achieves better results for every type of distortion. The worst correlation obtained by SSIM+ in the distortion type scenario is almost better than the best coefficient achieved by PSNR or SSIM.

For the HD resolution, TLabs Database Phase 1++ results for all sequences and con-

ditions show that MCE performs better than PSNR and SSIM. In this case no of the video quality metrics shows any type of dependency on the amount of scene cuts, camera motion or texture detail. In the "error-free" sequences both metrics, PSNR and SSIM, have very low performance. On the other hand if the type of content is known and the slicing concealment is used it would be recommended to use PSNR as prediction metric. In all other cases the advice would be to use the MCE video quality metric.

PSNR limitations have been also detailed. The variation of PSNR is a reliable indicator of the variation of quality for a specified codec and fixed content. For example, PSNR can be used for testing different codec optimisation strategies designed to maximise the subjective quality of a specified content. What is more, although the monotonic relationship between PSNR and subjective quality exists separately per content, it does not exist anymore across contents.

4.2 Possible future work

There are a number of possible extensions and applications for the work presented in this thesis. Some suggestions are as follows. First, continue with the evaluation of SSIM+ in other databases and with more metrics to compare with. Second, continue the research on new metrics or try to improve the existing ones. Finally, try to find no-reference metrics which can make good quality predictions.

5 References

References

[1] “The video Tsunami” [Eubanks, 71st IETF, March 08]

[2] http://www.itu.int/rec/T-REC-G.107

[3] S. Winkler and P.Mohandas. The evolution of video quality measurement: From psnr to hybrid metrics. IEEE Trans. Broadcasting, 54:1–9, 2008.

[4] BT.500-11, ITU-R Recommendation, ”Methodology for the subjective assessment of the quality of television pictures,” 2002.

[5] P.910, ITU-T Recommendation, ”Subjective video quality assesment methods for multimedia applications,” 2008.

[6] Tao Liu, ”Perceptual Quality Assessment of Videos Affected by Packet Losses”, January 2010

[7] S. Winkler, “Video quality and beyond,” in Proc. European Signal Processing Conference, Poznañ, Poland, September 3–7, 2007, invited paper.

[8] S. Winkler, “Perceptual video quality metrics—A review,” in Digital Video Image Quality and Perceptual Coding, H. R. Wu and K. R. Rao, Eds. Boca Raton, FL: CRC Press, 2005, ch. 5.

[9] Deepak S. Turaga, Yingwei Chen and Jorge Caviedes, “No Reference PSNR Estimation for Compressed Pictures,” Proceeding of ICIP, Vol. 3, PP. III61-III64, June 2002.

[10] A. Ichigaya, M. Kurozumi, N. Hara, Y. Nishida, and E. Nakasu, “A Method of Estimating Coding PSNR Using Quantized DCT Coefficients,” IEEE Transactions on Circuits and Systems for Video Technology, Volume 16, No. 12, pp.251-259, February 2006.

[11] Christian J. van den Branden Lambrecht, Olivier Verscheure, Chong Tean Ong and Vincent Darmstaedter, “Multimedia Decoder with Error Detection,” European Patent 1056297, 2000.

[12] http://www.itu.int/rec/R-REC-BT.601/en

[13] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh and Eero P. Simoncelli, ”Image Quality Assessment: From Error Visibility to Structural Similarity”

[14] Z. Wang, “Rate scalable Foveated image and video communications,” Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. Texas at Austin, Austin, TX, Dec. 2001.

[15] Z. Wang, A. C. Bovik, and L. Lu, “Why is image quality assessment so difficult,” in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 4, Orlando, FL, May 2002, pp. 3313–3316.

[16] Z. Wang, Rate scalable foveated image and video communications. PhD thesis, Dept. of ECE, The University of Texas at Austin, Dec. 2001.

[17] Z. Wang and A. C. Bovik, "A universal image quality index,"IEEE Signal Processing Letters, vol. 9, pp. 81–84, Mar. 2002. encoded images," Journal of Electronic Imaging, vol. 4, pp. 397– 406, Oct. 1995.

[18] W. S. Geisler and M. S. Banks, "Visual performance," in Handbook of Optics (M. Bass, ed.), McGraw-Hill, 1995.

[19] Z. Wang and A. C. Bovik, "Embedded foveation image coding," IEEE Trans. Image Processing, vol. 10, pp. 1397–1410, Oct. 2001.

[20] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," IEEE Trans. Pattern Analysis and Machine Intel ligence, vol. 22, pp. 970–982, Sept. 2000.

[21] U. Ra jashekar, L. K. Cormack, and A. C. Bovik, "Image features that draw fixations," in Proc. IEEE Int. Conf. Image Proc., vol. 3, pp. 313–316, Sept. 2003.

[22] Z. Wang, "The SSIM index for image quality assessment," http://www.cns.nyu.edu/ lcv/ssim/.

[23] Video Quality Experts Group (VQEG), "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, phase II," 2003 VQEG. Available: www.vqeg.org

[24] Margaret H Pinson and Stephen Wolf, "A New Standardized Method for Objectively Measuring Video Quality"

[25] S. Wolf and M. Pinson, "Video quality measurement techniques," NTIA Report 02-392, June 2002. Available: www.its.bldrdoc.gov/n3/video/documents.htm

[26] VMon and VQuad Results Description Manual, September 2008.

[27] Toru Yamada, Yoshihiro Miyamoto and Masahiro Serizawa, "No-Reference Video Quality

[28] Mitra Basu, Gaussian-Based Edge-Detection Methods—A Survey Estimation Based on Error-Concealment Effectiveness"

[29] K. Seshadrinathan, R. Soundararajan, A. C. Bovik and L. K. Cormack, "Study of Subjective and Objective Quality Assessment of Video", accepted for publication, IEEE Transactions on Image Processing, 2009.

[30] K. Seshadrinathan, R. Soundararajan, A. C. Bovik and L. K. Cormack, "A Subjective Study to Evaluate Video Quality Assessment Algorithms", to appear, SPIE Proceedings Human Vision and Electronic Imaging, Jan. 2010.

[31] http://live.ece.utexas.edu/research/quality/live_video.html

[32] http://www.itu.int/rec/T-REC-H.264/e

[33] D. A. Silverstein and J. E. Farrell, "The relationship between image fidelity and image quality," in Proc. IEEE Int. Conf. Image Proc., pp. 881–884, 1996.

[34] A. B. Watson, "Visual detection of spatial contrast patterns: Evaluation of five simple models," Optics Express, vol. 6, pp. 12–33, Jan. 2000.

[35] E. P. Simoncelli, “Statistical models for images: Compression, restoration and synthesis,” in Proc 31st Asilomar Conf on Signals, Systems and Computers, (Pacific Grove, CA), pp. 673–678, IEEE Computer Society, November 1997.

[36] J. Liu and P. Moulin, “Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients,” IEEE Trans. Image Processing, vol. 10, pp. 1647–1658, Nov. 2001.

[37] Zhou Wang and Alan C. Bovik, ”MSE: Love it or leave it. A new look at signal fidelity measures”

[38] Q. Huynh-Thu and M. Ghanbari ”Scope of validity of PSNR in image/video quality assessment” (2008)

A Annex

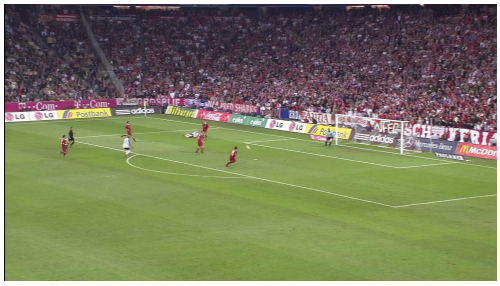
A.1 Databases characteristics

A.1.1 TLabs Database

In Figure 33 are shown the snapshots for all the TLabs Database contents.

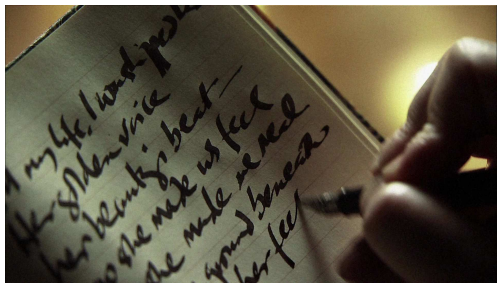
(a) Snapshot of content A.

(b) Snapshot of content B.



(c) Snapshot of content C.

(d) Snapshot of content D.



(e) Snapshot of content E.

Figure 33: Snapshots of TLabs Database.

In the tables shown below is are the characteristics of all the conditions used in TLabs Database.

Cond	Source Class	Codec	Keyframerate	Bitrate (Mbps)	Packet loss		Concealment	Decoder/Player
						Percentage	Type	
SD06-01	SD06-01	H.264	1fps	4	0	uniform		H.264 (Tsy)
SD06-02	SD06-02	H.264	1fps	2	0	uniform		H.264 (Tsy)
SD06-03	SD06-03	H.264	1fps	1	0	uniform		H.264 (Tsy)
SD07-07	SD06-01	H.264	1fps	4	0.06	uniform	frozen	H.264 (Tsy)
SD07-08	SD06-01	H.264	1fps	4	0.25	uniform	frozen	H.264 (Tsy)
SD07-09	SD06-01	H.264	1fps	4	0.125	uniform	frozen	H.264 (Tsy)
SD07-13	SD06-01	H.264	1fps	4	0.5	uniform	frozen	H.264 (Tsy)
SD10-06	SD06-01	H.264	1fps	4	0.125	uniform	slicing	H.264 (Tsy)
SD07-10	SD06-01	H.264	1fps	4	0.25	uniform	slicing	H.264 (Tsy)
SD07-11	SD06-01	H.264	1fps	4	0.5	uniform	slicing	H.264 (Tsy)
SD07-12	SD06-01	H.264	1fps	4	1	uniform	slicing	H.264 (Tsy)
SD10-05	SD06-01	H.264	1fps	4	4	uniform	slicing	H.264 (Tsy)
SD10-07	SD06-01	H.264	1fps	4	2	uniform	slicing	H.264 (Tsy)
SD08-01	SD06-02	H.264	1fps	2	0.06	uniform	frozen	H.264 (Tsy)
SD08-03	SD06-02	H.264	1fps	2	0.125	uniform	frozen	H.264 (Tsy)
SD08-04	SD06-02	H.264	1fps	2	0.125	uniform	slicing	H.264 (Tsy)
SD08-05	SD06-02	H.264	1fps	2	0.5	uniform	slicing	H.264 (Tsy)
SD08-06	SD06-02	H.264	1fps	2	2	uniform	slicing	H.264 (Tsy)
SD08-07	SD06-02	H.264	1fps	2	4	uniform	slicing	H.264 (Tsy)
SD08-08	SD06-02	H.264	1fps	2	0.25	uniform	frozen	H.264 (Tsy)
SD09-07	SD06-03	H.264	1fps	1	0.06	uniform	frozen	H.264 (Tsy)
SD09-09	SD06-03	H.264	1fps	1	0.125	uniform	frozen	H.264 (Tsy)
SD09-13	SD06-03	H.264	1fps	1	0.25	uniform	frozen	H.264 (Tsy)
SD10-12	SD06-03	H.264	1fps	1	0.125	uniform	slicing	H.264 (Tsy)
SD09-10	SD06-03	H.264	1fps	1	0.25	uniform	slicing	H.264 (Tsy)
SD09-11	SD06-03	H.264	1fps	1	0.5	uniform	slicing	H.264 (Tsy)
SD09-12	SD06-03	H.264	1fps	1	1	uniform	slicing	H.264 (Tsy)

Table 68: TLabs Phase 1+ SD video conditions I

Cond	Source Class	Codec	Keyframerate	Bitrate (Mbps)	Packet loss		Concealment	Decoder/Player	
						Percentage	Type		
SD10-13	SD06-03	H.264	1fps	1	2	uniform	slicing	H.264 (Tsy)	
SD10-11	SD06-03	H.264	1fps	1	4	uniform	slicing	H.264 (Tsy)	
SD11-01	SD11-1	uncompressed						Mplayer	
SD11-06	SD11-6	H.264	1fps	0,5	0	uniform		H.264 (Tsy)	
SD11-14	SD11-14	uncompressed			Dur: 20%	3x/0,8s/1,6s/0,8s	frozen	Mplayer	
SD11-18	SD11-18	uncompressed				blurring		Mplayer	
SD11-24	SD6-1	H.264	1fps	4	4	uniform	slicing	H.264 (Tsy)	
SD07-71		H.264	1fps	ph1	4	0.06	uniform	frozen	H.264 (Tsy)
SD07-81		H.264	1fps	ph1	4	0.25	uniform	frozen	H.264 (Tsy)
SD07-91		H.264	1fps	ph1	4	1	uniform	frozen	H.264 (Tsy)
SD10-61		H.264	1fps	ph1	4	0.125	uniform	slicing	H.264 (Tsy)
SD07-111		H.264	1fps	ph1	4	0.5	uniform	slicing	H.264 (Tsy)
SD10-51		H.264	1fps	ph1	4	4	uniform	slicing	H.264 (Tsy)

Table 69: TLabs Phase 1+ SD video conditions II

Cond	Source Class	Codec	Keyframerate	Bitrate (Mbps)	Packet loss		Concealment	Decoder/Player
						Percentage	Type	
HD01-01	HD01-01	MPEG2	1fps	32	0	uniform	slicing	Mplayer
HD06-01	HD06-01	H.264	1fps	16	0	uniform		H.264 (Tsy)
HD06-02	HD06-02	H.264	1fps	8	0	uniform		H.264 (Tsy)
HD06-03	HD6-3	H.264	1fps	4	0	uniform		H.264 (Tsy)
HD01-02	HD01-01	MPEG2	1fps	32	0.5	uniform	slicing	Mplayer
HD01-03	HD01-01	MPEG2	1fps	32	1	uniform	slicing	Mplayer
HD01-04	HD01-01	MPEG2	1fps	32	2	uniform	slicing	Mplayer
HD07-05	HD06-01	H.264	1fps	16	0.02	uniform	frozen	H.264 (Tsy)
HD07-07	HD06-01	H.264	1fps	16	0.06	uniform	frozen	H.264 (Tsy)
HD07-08	HD06-01	H.264	1fps	16	0.25	uniform	frozen	H.264 (Tsy)
HD07-09	HD06-01	H.264	1fps	16	0.125	uniform	frozen	H.264 (Tsy)
HD10-06	HD06-01	H.264	1fps	16	0.125	uniform	slicing	H.264 (Tsy)
HD07-10	HD06-01	H.264	1fps	16	0.25	uniform	slicing	H.264 (Tsy)
HD07-11	HD06-01	H.264	1fps	16	0.5	uniform	slicing	H.264 (Tsy)
HD07-12	HD06-01	H.264	1fps	16	1	uniform	slicing	H.264 (Tsy)
HD10-05	HD06-01	H.264	1fps	16	4	uniform	slicing	H.264 (Tsy)
HD10-07	HD06-01	H.264	1fps	16	2	uniform	slicing	H.264 (Tsy)
HD08-01	HD06-02	H.264	1fps	8	0.06	uniform	frozen	H.264 (Tsy)
HD08-02	HD06-02	H.264	1fps	8	0.02	uniform	frozen	H.264 (Tsy)
HD08-03	HD06-02	H.264	1fps	8	0.125	uniform	frozen	H.264 (Tsy)
HD08-04	HD06-02	H.264	1fps	8	0.125	uniform	slicing	H.264 (Tsy)
HD08-05	HD06-02	H.264	1fps	8	0.5	uniform	slicing	H.264 (Tsy)
HD08-06	HD06-02	H.264	1fps	8	2	uniform	slicing	H.264 (Tsy)
HD09-07	HD06-03	H.264	1fps	4	0.06	uniform	frozen	H.264 (Tsy)
HD09-08	HD06-03	H.264	1fps	4	0.02	uniform	frozen	H.264 (Tsy)
HD09-09	HD06-03	H.264	1fps	4	0.125	uniform	frozen	H.264 (Tsy)
HD10-12	HD06-03	H.264	1fps	4	0.125	uniform	slicing	H.264 (Tsy)
HD09-10	HD06-03	H.264	1fps	4	0.25	uniform	slicing	H.264 (Tsy)
HD09-11	HD06-03	H.264	1fps	4	0.5	uniform	slicing	H.264 (Tsy)
HD09-12	HD06-03	H.264	1fps	4	1	uniform	slicing	H.264 (Tsy)

Table 70: TLabs Phase 1+ HD video conditions I

Cond	Source Class	Codec	Keyframerate	Bitrate (Mbps)	Packet loss		Concealment	Decoder/Player
						Percentage	Type	
HD10-13	HD06-03	H.264	1fps	4	2	uniform	slicing	H.264 (Tsy)
HD11-01	HD11-01	uncompressed						Mplayer
HD11-06	HD11-06	H.264	1fps	2	0	uniform		H.264 (Tsy)
HD11-14	HD11-14	uncompressed			Dur: 20%	3x/0,8s/1,6s/0,8s	frozen	Mplayer
HD11-18	HD11-18	uncompressed				medium blurring		Mplayer
HD11-24	HD06-01	H.264	1fps	16	4	uniform	slicing	H.264 (Tsy)

Table 71: TLabs Phase 1+ HD video conditions II

Cond	Source Class	Codec	Keyframerate	Bitrate (Mbps)	Packet loss		Concealment	Decoder/Player
						Percentage	Type	
SD06-01	SD06-01	H.264	1fps	4	0	uniform		H.264 (Tsy)
SD06-02	SD06-02	H.264	1fps	2	0	uniform		H.264 (Tsy)
SD06-03	SD06-03	H.264	1fps	1	0	uniform		H.264 (Tsy)
SD07-07	SD06-01	H.264	1fps	4	0.06	uniform	frozen	H.264 (Tsy)
SD07-08	SD06-01	H.264	1fps	4	0.25	uniform	frozen	H.264 (Tsy)
SD07-09	SD06-01	H.264	1fps	4	0.125	uniform	frozen	H.264 (Tsy)
SD07-13	SD06-01	H.264	1fps	4	0.5	uniform	frozen	H.264 (Tsy)
SD10-06	SD06-01	H.264	1fps	4	0.125	uniform	slicing	H.264 (Tsy)
SD07-10	SD06-01	H.264	1fps	4	0.25	uniform	slicing	H.264 (Tsy)
SD07-11	SD06-01	H.264	1fps	4	0.5	uniform	slicing	H.264 (Tsy)
SD07-12	SD06-01	H.264	1fps	4	1	uniform	slicing	H.264 (Tsy)
SD10-05	SD06-01	H.264	1fps	4	4	uniform	slicing	H.264 (Tsy)
SD10-07	SD06-01	H.264	1fps	4	2	uniform	slicing	H.264 (Tsy)
SD08-01	SD06-02	H.264	1fps	2	0.06	uniform	frozen	H.264 (Tsy)
SD08-03	SD06-02	H.264	1fps	2	0.125	uniform	frozen	H.264 (Tsy)
SD08-04	SD06-02	H.264	1fps	2	0.125	uniform	slicing	H.264 (Tsy)
SD08-05	SD06-02	H.264	1fps	2	0.5	uniform	slicing	H.264 (Tsy)
SD08-06	SD06-02	H.264	1fps	2	2	uniform	slicing	H.264 (Tsy)
SD08-08	SD06-02	H.264	1fps	2	0.25	uniform	frozen	H.264 (Tsy)

Table 72: TLabs Phase 1++ SD video conditions I

Cond	Source Class	Codec	Keyframerate	Bitrate (Mbps)	Packet loss		Concealment	Decoder/Player
						Percentage	Type	
SD09-07	SD06-03	H.264	1fps	1	0.06	uniform	frozen	H.264 (Tsy)
SD09-09	SD06-03	H.264	1fps	1	0.125	uniform	frozen	H.264 (Tsy)
SD09-13	SD06-03	H.264	1fps	1	0.25	uniform	frozen	H.264 (Tsy)
SD09-14	SD06-03	H.264	1fps	1	0.5	uniform	frozen	H.264 (Tsy)
SD10-12	SD06-03	H.264	1fps	1	0.125	uniform	slicing	H.264 (Tsy)
SD09-10	SD06-03	H.264	1fps	1	0.25	uniform	slicing	H.264 (Tsy)
SD09-11	SD06-03	H.264	1fps	1	0.5	uniform	slicing	H.264 (Tsy)
SD09-12	SD06-03	H.264	1fps	1	1	uniform	slicing	H.264 (Tsy)
SD10-13	SD06-03	H.264	1fps	1	2	uniform	slicing	H.264 (Tsy)
SD10-11	SD06-03	H.264	1fps	1	4	uniform	slicing	H.264 (Tsy)
SD11-01	SD11-1	uncompressed						Mplayer
SD11-06	SD11-6	H.264	1fps	0,5	0	uniform		H.264 (Tsy)
SD11-14	SD11-14	uncompressed			Dur: 20%	3x/0,8s/1,6s/0,8s	frozen	Mplayer
SD11-18	SD11-18	uncompressed				medium blurring		Mplayer
SD11-24	SD6-1	H.264	1fps	4	4	uniform	slicing	H.264 (Tsy)

Table 73: TLabs Phase 1++ SD video conditions II

Cond	Source Class	Codec	Keyframerate	Bitrate (Mbps)	Packet loss		Concealment	Decoder/Player
						Percentage	Type	
HD06-01	HD06-01	H.264	1fps	16	0	uniform		H.264 (Tsy)
HD06-02	HD06-02	H.264	1fps	8	0	uniform		H.264 (Tsy)
HD06-03	HD6-3	H.264	1fps	4	0	uniform		H.264 (Tsy)
HD07-05	HD06-01	H.264	1fps	16	0.02	uniform	frozen	H.264 (Tsy)
HD07-07	HD06-01	H.264	1fps	16	0.06	uniform	frozen	H.264 (Tsy)
HD07-08	HD06-01	H.264	1fps	16	0.25	uniform	frozen	H.264 (Tsy)
HD07-09	HD06-01	H.264	1fps	16	0.125	uniform	frozen	H.264 (Tsy)
HD10-06	HD06-01	H.264	1fps	16	0.125	uniform	slicing	H.264 (Tsy)
HD07-10	HD06-01	H.264	1fps	16	0.25	uniform	slicing	H.264 (Tsy)
HD07-11	HD06-01	H.264	1fps	16	0.5	uniform	slicing	H.264 (Tsy)
HD07-12	HD06-01	H.264	1fps	16	1	uniform	slicing	H.264 (Tsy)
HD10-05	HD06-01	H.264	1fps	16	4	uniform	slicing	H.264 (Tsy)
HD10-07	HD06-01	H.264	1fps	16	2	uniform	slicing	H.264 (Tsy)
HD08-01	HD06-02	H.264	1fps	8	0.06	uniform	frozen	H.264 (Tsy)
HD08-02	HD06-02	H.264	1fps	8	0.02	uniform	frozen	H.264 (Tsy)
HD08-03	HD06-02	H.264	1fps	8	0.125	uniform	frozen	H.264 (Tsy)
HD08-04	HD06-02	H.264	1fps	8	0.125	uniform	slicing	H.264 (Tsy)
HD08-05	HD06-02	H.264	1fps	8	0.5	uniform	slicing	H.264 (Tsy)
HD08-06	HD06-02	H.264	1fps	8	2	uniform	slicing	H.264 (Tsy)

Table 74: TLabs Phase 1++ HD video conditions I

Cond	Source Class	Codec	Keyframerate	Bitrate (Mbps)	Packet loss		Concealment	Decoder/Player
						Percentage	Type	
HD09-07	HD06-03	H.264	1fps	4	0.06	uniform	frozen	H.264 (Tsy)
HD09-08	HD06-03	H.264	1fps	4	0.02	uniform	frozen	H.264 (Tsy)
HD09-09	HD06-03	H.264	1fps	4	0.125	uniform	frozen	H.264 (Tsy)
HD09-13	HD06-03	H.264	1fps	4	0.25	uniform	frozen	H.264 (Tsy)
HD10-12	HD06-03	H.264	1fps	4	0.125	uniform	slicing	H.264 (Tsy)
HD09-10	HD06-03	H.264	1fps	4	0.25	uniform	slicing	H.264 (Tsy)
HD09-11	HD06-03	H.264	1fps	4	0.5	uniform	slicing	H.264 (Tsy)
HD09-12	HD06-03	H.264	1fps	4	1	uniform	slicing	H.264 (Tsy)
HD10-13	HD06-03	H.264	1fps	4	2	uniform	slicing	H.264 (Tsy)
HD10-11	HD06-03	H.264	1fps	4	4	uniform	slicing	H.264 (Tsy)
HD11-01	HD11-01	uncompressed						Mplayer
HD11-06	HD11-06	H.264	1fps	2	0	uniform		H.264 (Tsy)
HD11-14	HD11-14	uncompressed			Dur: 20%	3x/0,8s/1,6s/0,8s	frozen	Mplayer
HD11-18	HD11-18	uncompressed				medium blurring		Mplayer
HD11-24	HD06-01	H.264	1fps	16	4	uniform	slicing	H.264 (Tsy)

Table 75: TLabs Phase 1++ HD video conditions II

A.1.2 LIVE Video Quality Database

In Figure 34 are shown the snapshots for all the LIVE Database contents.

(a) Snapshot of the blue sky reference sequence.

(b) Snapshot of the mobile and calendar reference sequence.

(c) Snapshot of the pedestrian area reference sequence.

(d) Snapshot of the park run reference sequence.

(e) Snapshot of the riverbed reference sequence.

(f) Snapshot of the rushhour reference sequence.

(g) Snapshot of the sunflower reference sequence.

(h) Snapshot of the shields reference sequence.

(i) Snapshot of the station reference sequence.

(j) Snapshot of the tractor reference sequence.

Figure 34: Snapshots of LIVE Database.

A.1.3 VQM results

In tables 77, 78, 76 and 79 are shown the results for the firsts 15 seconds. In tables 81, 82, 80 and 83 for the last 15 seconds. As can be seen the Pearson correlation values for both methodologies are approximately the same.

- Phase 1++ 0-15:

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.9998	1.0000	0.6667	1.5802
B	0.9769	1.0000	0.3333	1.4445
C	0.9998	1.0000	1.0000	2.6582
D	0.9840	1.0000	0.3333	1.4656
E	0.9771	1.0000	1.0000	2.8132
ALL	0.6893	0.7179	0.6667	2.0838

Table 76: VQM results for the first 15 seconds of TLabs Database Phase 1++ for the "error-free" scenario.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	-0.2621	-0.2545	0.6364	3.9230
B	0.2389	0.4182	0.6364	2.9930
C	0.0512	0.0909	1.0000	4.4290
D	-0.1112	-0.1364	0.9091	3.7416
E	-0.3888	-0.2182	0.9091	4.6727
ALL	-0.0268	-0.0163	0.7818	3.9949

Table 77: VQM results for the first 15 seconds of TLabs Database Phase 1++ for the freezing scenario.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.5595	0.5719	0.7333	3.0785
B	0.7615	0.7857	0.6000	2.2754
C	0.8551	0.8250	1.0000	3.2729
D	0.7914	0.7964	0.6000	2.8764
E	0.7878	0.7321	1.0000	4.2846
ALL	0.6471	0.6428	0.7867	3.2248

Table 78: VQM results for the first 15 seconds of TLabs Database Phase 1++ for the slicing scenario.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.2909	0.3227	0.5517	3.3163
B	0.6228	0.6626	0.4138	2.5083
C	0.5259	0.4975	0.9655	3.7030
D	0.5285	0.5340	0.5172	3.1324
E	0.4751	0.4182	0.9655	4.3123
ALL	0.4366	0.4191	0.7034	3.4470

Table 79: VQM results for the first 15 seconds of TLabs Database Phase 1++ for all contents and conditions.

- Phase 1++ 1-16:

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.9995	1.0000	0.6667	1.5664
B	0.9781	1.0000	0.3333	1.4603
C	0.9999	1.0000	1.0000	2.6837
D	0.9850	1.0000	0.3333	1.5207
E	0.9796	1.0000	1.0000	2.7767
ALL	0.6954	0.7179	0.6667	2.0886

Table 80: VQM results for the last 15 seconds of TLabs Database Phase 1++ for the "error-free" scenario.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	-0.2616	-0.3000	0.6364	3.9092
B	0.2438	0.4545	0.6364	3.0091
C	0.0613	0.0909	1.0000	4.4558
D	-0.1063	-0.1364	0.9091	3.7993
E	-0.3748	-0.2273	0.9091	4.6265
ALL	-0.0298	-0.0294	0.8000	4.0008

Table 81: VQM results for the last 15 seconds of TLabs Database Phase 1++ for the freezing scenario.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.5683	0.5719	0.7333	3.0558
B	0.7717	0.8000	0.6000	2.2515
C	0.8623	0.8286	1.0000	3.2857
D	0.8320	0.8714	0.6000	2.8899
E	0.8094	0.7750	1.0000	4.1313
ALL	0.6757	0.6768	0.8000	3.1819

Table 82: VQM results for the last 15 seconds of TLabs Database Phase 1++ for the slicing scenario.

Content	Pearson corr	Spearman corr	Outlier rat.	RMSE
A	0.2954	0.3197	0.5517	3.2985
B	0.6308	0.6704	0.4138	2.5054
C	0.5316	0.4995	0.9655	3.7229
D	0.5548	0.5650	0.5172	3.1677
E	0.4982	0.4414	0.9655	4.2124
ALL	0.4544	0.4305	0.6966	3.4292

Table 83: VQM results for the last 15 seconds of TLabs Database Phase 1++ for all contents and conditions.